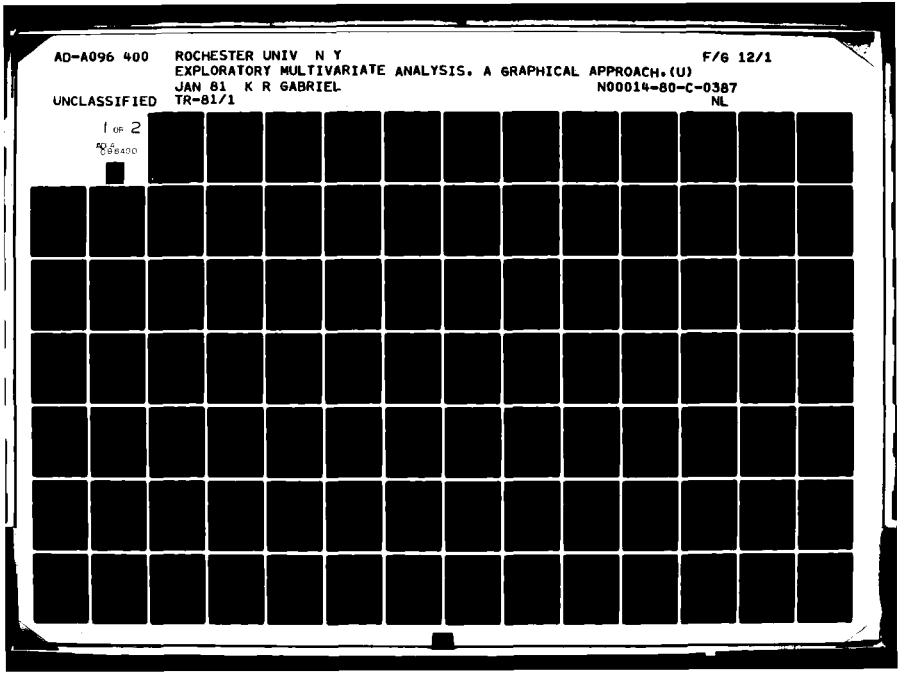

AD-A096 400
ROCHESTER UNIV N Y
F/G 12/1
EXPLORATORY MULTIVARIATE ANALYSIS. A GRAPHICAL APPROACH.(U)
JAN 81 K R GABRIEL
N00014-80-C-0387
UNCLASSIFIED
TR-81/1
NL
1 OF 2
AD A096400

LEVEL II (12)

AD A096400

EXPLORATORY MULTIVARIATE ANALYSIS

-- A GRAPHICAL APPROACH

BY

K. R. Gabriel

Department of Statistics
and
Division of Biostatistics

TECHNICAL REPORT 81/1

University of Rochester

Rochester, New York 14627 USA

JAN. 1981

This work was supported in part by ONR Contract N00014-80-C-0387

EXPLORATORY MULTIVARIATE ANALYSIS

-- A GRAPHICAL APPROACH

BY

K. R. Gabriel

Department of Statistics
and
Division of Biostatistics

TECHNICAL REPORT 81/1

University of Rochester

Rochester, New York 14627 USA

To Appear In

Probability, Statistics and

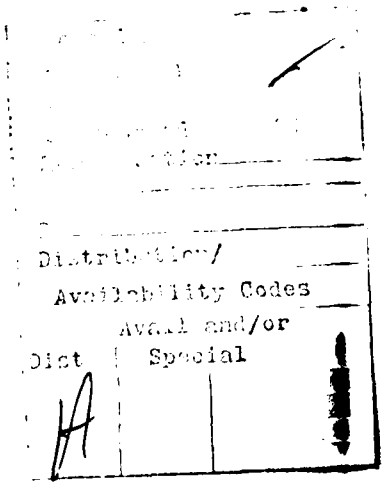
Decision Making in Meteorology

Allan H. Murphy and Richard W. Katz, Eds.

This work was supported in part by ONR Contract N00014-80-C-0387

INTRODUCTION

This Chapter presents a very idiosyncratic view of multivariate analysis and reflects what the author has found useful in his statistical practice. It stresses exploration, virtually ignores tests of significance, and emphasizes a particular graphical technique developed by the author -- biplot display. The author hopes that the Chapter will help its readers towards a better grasp of the structure of multivariate data and the fundamentals of multivariate analysis. He hopes that, despite the Chapter's personal bias, it will also help readers who will wish to pursue multivariate analysis in its more classical form, for which references are given in this literature.



1. ONE BATCH OF MULTIVARIATE DATA AND THEIR DESCRIPTIVE STATISTICS

1.1. A multivariate data matrix

The essence of multivariability is that several variables are observed on each unit. Thus, if the units are days, one might observe maximum and minimum temperatures on each day, as well as precipitation and surface biometric pressure at 6 a.m., 12 noon, 6 p.m. and midnight; these would be 7-variate observations. It is convenient to think of the data as a matrix $Z_{(nxm)}$ in which row $\underline{z}'_i$ $(i=1,\ldots,n)$ contains the m-variate observations for unit i -- out of n units -- and each column $\underline{z}_{(v)}$ $(v=1,\ldots,m)$ contains all n units' observations on the v-th variable and $z_{i,v}$ is unit i's observation on variable v. Thus, in this example, element $z_{2,3}$ would be the total precipitation on day 2, $\underline{z}'_2$ would be the seven variate observations on day 2 and $\underline{z}_{(3)}$ the n days' observations on the third variable -- total precipitation.

We adopt the convention of denoting a matrix by a Latin capital letter and any of its elements by the corresponding lower case letter with two indices, which indicate, respectively, the row and column in which the element is located. We denote both rows and columns of the matrix by the lower case letter underlined and with a single index: if the index is in parentheses, a column is denoted; if no parentheses are shown, a row is indicated.

For a detailed illustration, consider the data of Table 1: mean monthly temperatures for 20 stations during 6 months of the year 1951. Here, n=20, m=6, and $z_{3,1}$ is the third station's mean January temperature, whereas $z_{1,3}$ is the first station's mean May temperature. The location of the 20 stations is shown on the map of Figure 1.

A first glance at the data matrix like this is apt to be somewhat confusing. Some idea of the general pattern can be obtained from the means and standard deviations -- shown at the bottom of Table 1. The temperature averages are seen to be much the same in all the six months (evidently because the stations are spread on both sides of the equator). However, there is considerable variation from station to station, as evidenced by the large standard deviations: these are around 50 (i.e., 5 degrees centigrade) for each month, though a bit less for March and May.

TABLE 1

Mean Monthly Temperatures at Certain American Stations, 1951

(10 x centigrade)

Station (See Fig 1) i \ v	1 JAN	2 MAR	3 MAY	4 JUL	5 SEP	6 NOV
1	260	251	224	196	215	240
2	259	268	249	226	275	268
3	325	226	193	163	184	234
4	204	188	149	129	157	205
5	192	172	156	147	145	172
6	269	275	254	233	247	263
7	273	278	263	260	273	277
8	255	256	253	250	267	271
9	248	251	247	242	254	260
10	259	256	259	236	268	261
11	124	136	132	130	136	127
12	241	240	216	191	186	191
13	247	256	253	252	266	265
14	261	281	277	250	276	278
15	263	269	271	271	269	268
16	242	249	273	284	280	270
17	192	218	266	278	273	215
18	198	204	237	278	284	254
19	132	174	238	283	272	146
20	68	125	228	298	255	106
Means	225.60	228.65	231.90	230.00	239.10	228.55
Standard Deviations	59.08	45.93	41.56	51.42	48.04	52.10

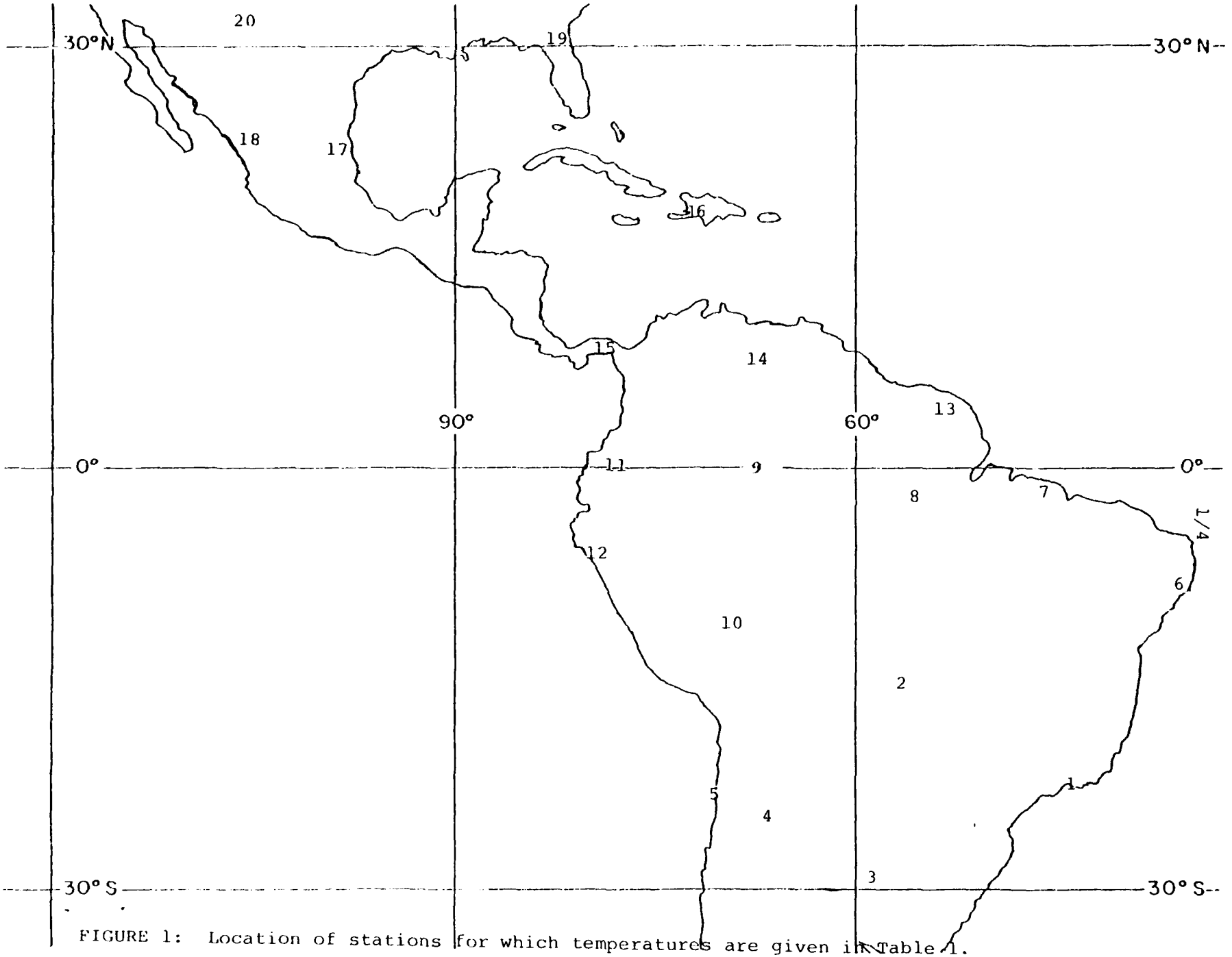


FIGURE 1: Location of stations for which temperatures are given in Table 1.

1.2. Summary statistics for the variables' configuration

The common statistics for the batch of n units are readily obtained from matrix Z. Thus, the means $\bar{z}_{(1)},\ldots,\bar{z}_{(m)}$ of all m variables are arrayed in vector

$$\underline{\bar{z}}' = (1/n)\ \underline{1}_n'Z, \tag{1.1}$$

where $\underline{1}_n$ is a vector of n ones. Deviations from each variable's mean are given in matrix

$$Y = Z - \underline{1}_n\underline{\bar{z}}' , \tag{1.2}$$

which has typical element

$$y_{i,v} = z_{i,v} - \bar{z}_{(v)} . \tag{1.3}$$

The means for the temperature illustration were noted in Table 1, and the deviations $y_{i,v}$ from the means are shown in Table 2. Thus, $y_{1,3} = 7.9 = 224.0 - 231.9 = z_{1,3} - \bar{z}_{(3)}$.

From these one may compute the variance matrix (often referred to as the variance - covariance matrix)

$$S = \frac{1}{n}\ Y'Y, \tag{1.4}$$

the standard deviations

$$s_v = \sqrt{s_{v,v}} \qquad (v=1,\ldots,m), \tag{1.5}$$

and, defining the diagonal matrix with elements $d_{v,v} = s_v$ as D_s, the correlation matrix

$$R = D_s^{-1}SD_s^{-1}, \tag{1.6}$$

whose elements are the correlations $r_{v,v'}$ between variables.

For the temperature illustration, the standard deviations were shown in Table 1, and the variance and correlation matrices are given in Tables 3 and 4, respectively.

TABLE 2

Deviations from Monthly Means - Data of Table 1

i \ v	1 JAN	2 MAR	3 MAY	4 JUL	5 SEP	6 NOV
1	34.4	22.35	-7.9	-34	-24.1	11.45
2	33.4	39.35	17.1	-4	35.9	39.45
3	99.4	-2.65	-38.9	-64	-55.1	5.45
4	-21.6	-40.65	-82.9	-101	-82.1	-23.55
5	-33.6	-56.65	-75.9	-83	-94.1	-56.55
6	43.4	46.35	22.1	3	7.9	34.45
7	47.4	49.35	31.1	30	33.9	48.45
8	29.4	27.35	21.1	20	27.9	42.45
9	22.4	22.35	15.1	12	14.9	31.45
10	33.4	27.35	27.1	6	28.9	32.45
11	-101.6	-92.65	-99.9	-100	-103.1	-101.55
12	15.4	11.35	-15.9	-39	-53.1	-37.55
13	21.4	27.35	21.1	22	26.9	36.45
14	35.4	52.35	45.1	20	36.9	49.45
15	37.4	40.35	39.1	41	29.9	39.45
16	16.4	20.35	41.1	54	40.9	41.45
17	-33.6	-10.65	34.1	48	33.9	-13.55
18	-27.6	-24.65	5.1	48	44.9	25.45
19	-93.6	-54.65	6.1	53	32.9	-82.55
20	-157.6	-103.65	-3.9	68	15.9	-122.55

TABLE 3

Variance Matrix of Mean Temperatures - Data of Table 1

v	v' 1 JAN	2 MAR	3 MAY	4 JUL	5 SEP	6 NOV
1 JAN	3490.7	2376.4	895.86	-277.15	463.09	2635.4
2 MAR	2376.4	2109.5	1324.7	612.2	1117.8	2221.4
3 MAY	895.86	1324.7	1726.8	1843.3	1873.6	1388.2
4 JUL	-277.15	612.2	1843.3	2644.5	2300.4	714.15
5 SEP	463.09	1117.8	1873.6	2300.4	2307.5	1338.1
6 NOV	2635.4	2221.4	1388.2	714.15	1338.1	2714.1

TABLE 4

Correlation Matrix of Mean Temperatures -
Data of Table 1

v	v' 1 JAN	2 MAR	3 MAY	4 JUL	5 SEP	6 NOV
1 JAN	1	0.87573	0.36489	-0.091219	0.16317	0.8562
2 MAR	0.87573	1	0.69405	0.2592	0.50664	0.92836
3 MAY	0.36489	0.69405	1	0.86259	0.93859	0.64123
4 JUL	-0.091219	0.2592	0.86259	1	0.93124	0.26656
5 SEP	0.16317	0.50664	0.93859	0.93124	1	0.53471
6 NOV	0.8562	0.92836	0.64123	0.26656	0.53471	1

The configuration of correlations in an (mxm) matrix R may not be easy to grasp at first, especially if m is 10 or more. In the present illustration with m=6, one may begin to study Table 4 by concentrating on the highest correlations. One notices two distinct sheaves of months: November, January and March are highly intercorrelated and so are, even more strongly, May, July and September. The correlations between months not belonging to the same sheaf (or season) are seen to be much lower.

Such a perusal of correlations is not always easy, especially if the variables do not group neatly into highly inter-correlated sheaves. It is sometimes helpful also to consider the inverse of the variance matrix, i.e.,

$$S^{-1} = (\frac{1}{n} Y'Y)^{-1}, \qquad (1.7)$$

because its elements $s^{v,v'}$ have the following interpretation in terms of the multiple regression coefficients of the v-th variable on <u>all</u> <u>other</u> variables. Take the v-th row of S^{-1}, divide each off-diagonal element by the diagonal element and change sign - then

$$b_{v,v'} = s^{v,v'}/s^{v,v} \qquad (1.8)$$

is the coefficient of variable v' in the regression for variable v. Furthermore, using diagonal terms from both S and S^{-1}

$$r_v^2 = 1 - (s_{v,v} s^{v,v})^{-1} \qquad (1.9)$$

gives the multiple correlation of variable v on all m-1 other variables.

For the temperature illustration, the inverse S^{-1} and the multiple correlation and regression coefficients are

given in Tables 5 and 6, respectively. Each month's temperature is seen to be pretty highly correlated with the temperatures of all other months. And of course the multiple correlations are all higher than the correlations with individual variables, that is,

$$r_v \geq r_{v,v'} \quad (v' \neq v). \tag{1.10}$$

It is interesting to see the pattern of regression coefficients: Each month's coefficients with adjacent months are positive but with months about half a year away (2 or 3 variables away in the circular order ...123456123...) the coefficients are negative. This makes good sense in terms of consistent seasonal patterns.

TABLE 5

Inverse of Variance Matrix S

	1	2	3	4	5	6
1	0.003122	-0.0019469	-0.00089546	0.000063023	0.0021828	-0.0020728
2	-0.0019469	0.019515	-0.024644	0.011706	0.002688	-0.0058826
3	-0.00089546	-0.024644	0.042981	-0.019968	-0.0075252	0.0080195
4	0.000063023	0.011706	-0.019968	0.015857	-0.0044671	-0.0013988
5	0.0021828	0.002688	-0.0075252	-0.0044671	0.012391	-0.005404
6	-0.0020728	-0.0058826	0.0080195	-0.0013988	-0.005404	0.0061264

TABLE 6

Multiple Correlation and Regressions of Each Variable on All Others (Correlations in Diagonal, Regression Coefficients Off-Diagonal)

Regression of Variable	onto variable v					
	1	2	3	4	5	6
1	.9530	.6236	.2868	-.0202	-.6992	.6639
2	.0998	.9878	1.2628	-.5998	-.1377	.3014
3	.0208	.5734	.9932	.4646	.1751	-.1866
4	-.0040	-.7382	1.2593	.9880	.2817	.0882
5	-.1762	-.2169	.6073	.3605	.9824	.4361
6	.3383	.9602	-1.3090	.2283	.8821	.9695

1.3. Distances in the units' scatter

The above statistics describe the configuration of the variables (monthly temperatures) for the entire batch of units -- no attention being paid to the individual units (stations). If one is interested in the individual units, their similarities and differences, one needs a decription of the scatter of the units. For this purpose one would calculate the units' metric

$$U_{(nxn)} = YS^{-1}Y' \quad , \tag{1.11}$$

which can be interpreted in terms of standardized distances as follows: The diagonal elements of U

$$\begin{aligned} u_{i,i} &= \underline{y}_i' \, S^{-1} \, \underline{y}_i \\ &= (\underline{z}_i - \bar{\underline{z}})' S^{-1} \, (\underline{z}_i - \bar{\underline{z}}) \quad , \end{aligned} \tag{1.12}$$

are squares of standardized distances of units i from the centroid, i.e., from the multivariate mean of the batch. The tetrad differences

$$\begin{aligned} d_{i,e} &= (u_{i,i} - u_{i,e} - u_{e,i} + u_{e,e}) \\ &= (\underline{y}_i - \underline{y}_e)' S^{-1} (\underline{y}_i - \underline{y}_e) \\ &= (\underline{z}_i - \underline{z}_e)' S^{-1} (\underline{z}_i - \underline{z}_e) , \end{aligned} \tag{1.13}$$

are squares of standardized distances between units i and e. Such distances should be understood as measuring statistical differences simultaneously on all m variables. They are equal to zero if and only if the units compared have equal observations on all variables, and they increase when the differences in any one or more variable becomes larger.

TABLE 7

Inter-station Standardized Distance - Data of Table 1

Station	Station																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.0	3.0	4.2	2.7	2.3	1.1	2.4	1.7	1.4	2.6	2.9	2.3	1.7	2.1	2.1	2.8	2.8	4.0	3.7	3.9
2	3.0	0.0	5.3	3.6	4.7	3.3	3.4	3.0	3.3	2.9	4.1	4.7	3.0	3.3	4.2	4.7	4.3	4.4	3.4	5.1
3	4.2	5.3	0.0	5.1	4.4	4.9	5.0	4.4	4.5	4.5	5.6	5.0	4.7	5.4	4.7	4.9	5.0	5.2	5.2	5.6
4	2.7	3.6	5.1	0.0	2.2	3.2	3.7	2.7	2.7	3.8	1.8	4.3	2.8	3.8	3.8	3.9	4.2	3.5	4.6	4.7
5	2.3	4.7	4.4	2.2	0.0	2.9	3.7	2.8	2.5	3.9	2.1	3.3	2.9	3.7	3.0	3.0	3.4	4.0	4.6	3.8
6	1.1	3.3	4.9	3.2	2.9	0.0	1.7	1.7	1.3	3.3	3.6	2.2	1.4	2.4	1.4	2.7	3.1	4.0	3.8	4.0
7	2.4	3.4	5.0	3.7	3.7	1.7	0.0	1.9	2.0	4.3	4.3	3.1	1.5	3.7	1.8	3.3	4.3	3.4	3.5	4.4
8	1.7	3.0	4.4	2.7	2.8	1.7	1.9	0.0	0.6	2.8	3.5	3.6	0.6	2.4	1.7	2.0	2.8	2.5	3.6	3.7
9	1.4	3.3	4.5	2.7	2.5	1.3	2.0	0.6	0.0	2.8	3.3	3.2	0.6	2.3	1.4	1.8	2.6	2.9	3.7	3.6
10	2.6	2.9	4.5	3.8	3.9	3.3	4.3	2.8	2.8	0.0	4.0	4.5	3.1	1.8	3.8	3.3	2.1	4.5	4.2	4.4
11	2.9	4.1	5.6	1.8	2.1	3.6	4.3	3.5	3.3	4.0	0.0	4.0	3.4	3.8	4.2	4.3	3.9	4.4	4.2	3.9
12	2.3	4.7	5.0	4.3	3.3	2.2	3.1	3.6	3.2	4.5	4.0	0.0	3.3	3.9	2.8	4.2	4.1	5.7	4.1	4.3
13	1.7	3.0	4.7	2.8	2.9	1.4	1.5	0.6	0.6	3.1	3.4	3.3	0.0	2.6	1.5	2.2	3.1	2.7	3.5	3.7
14	2.1	3.3	5.4	3.8	3.7	2.4	3.7	2.4	2.3	1.8	3.8	3.9	2.6	0.0	3.0	2.8	2.0	4.6	4.5	4.4
15	2.1	4.2	4.7	3.8	3.0	1.4	1.8	1.7	1.4	3.8	4.2	2.8	1.5	3.0	0.0	1.8	3.1	3.5	4.0	3.7
16	2.8	4.7	4.9	3.9	3.0	2.7	3.3	2.0	1.8	3.3	4.3	4.2	2.2	2.8	1.8	0.0	2.2	3.1	4.7	3.6
17	2.8	4.3	5.0	4.2	3.4	3.1	4.3	2.8	2.6	2.1	3.9	4.1	3.1	2.0	3.1	2.2	0.0	4.3	4.1	2.9
18	4.0	4.4	5.2	3.5	4.0	4.0	3.4	2.5	2.9	4.5	4.4	5.7	2.7	4.6	3.5	3.1	4.3	0.0	4.4	4.2
19	3.7	3.4	5.2	4.6	4.6	3.8	3.5	3.6	3.7	4.2	4.2	4.1	3.5	4.5	4.0	4.7	4.1	4.4	0.0	2.8
20	3.9	5.1	5.6	4.7	3.8	4.0	4.4	3.7	3.6	4.4	3.9	4.3	3.7	4.4	3.7	3.6	2.9	4.2	2.8	0.0

The matrix of standardized distances $\sqrt{d}_{i,e}$ between each pair of stations on the six month's temperature data is given in Table 7. All distances are positive except the "self-distances" in the diagonal which are identically zero; the matrix is symmetric, that is, the i-to-e distance $\sqrt{d}_{i,e}$ equals the e-to-i distance $\sqrt{d}_{e,i}$. Small distances, such as $d_{8,9} = 0.6$, indicate that stations 8 and 9 have very similar mean monthly temperatures; whereas large distances, such as $d_{2,3} = 5.3$, show that very considerable differences in mean monthly temperatures exist between stations 2 and 3. (The reader can verify this from Table 1.)

It is difficult to inspect a table of distances of this magnitude (not to speak of distance matrices for a hundred or more units). We shall therefore require methods of disentangling the pattern of distances of a scatter of units and of making some sense of such a distance matrix -- these will be discussed below in Section 4.

1.4. Some further remarks on standardized statistical distances

To understand the u and d statistics it is well to begin by considering standardized difference between units i and e on one particular variable. Thus on the v-th variable alone the distance would be

$$\sqrt{d}_{i,e(v)} = |y_{i,v} - y_{e,v}|/\sqrt{s}_{v,v} \ . \qquad (1.14)$$

Similarly, for the linear combination of variables (LCV) with coefficients $\underline{a} = (a_1,\ldots,a_m)'$ the standardized difference would be

$$\sqrt{d}_{i,e(\underline{a})} = |\underline{a}'\underline{y}_i - \underline{a}'\underline{y}_e|/\sqrt{\underline{a}'S\underline{a}} \ , \qquad (1.15)$$

since the variance of that LCV is $\underline{a}'S\underline{a}$. A generalized i-to-e distance, for all variables and LCVs together, can then reasonably be defined as the maximum of all such LCVs' differences. But it can be proved that this maximum satisfies

$$\max_{a_1,\ldots,a_m} \sqrt{d}_{i,e(\underline{a})} = \sqrt{d}_{i,e} \ , \qquad (1.16)$$

so that the proposed generalized i-to-e distance of (1.13) can be regarded as a maximum difference over all variables and LCVs.

A similar explanation can be given for the structure of the standardized distance $\sqrt{u}_{i,i}$ to the centroid.

1.5. What are the units and what are the variables?

Statistics textbooks usually treat only the description of the variables' configurations and ignore that of the units' scatter. This is presumably because statisticians have mostly been concerned with random samples in which the individual units are of no interest in themselves. In practical data analysis, however, the units are often of real interest and their description is as relevant as that of the variables. Indeed, it is not always obvious which of the classifications of data one wants to regard as units and which as variables. In the example mentioned above, time has appeared as a variable - but in a series of successive observations on given stations or measurements it might appear as a unit.

In any particular application, the decision of what to regard as units and what as variables will determine what will be weighted equally and what will be standardized. In the analyses discussed above, the treatment of the rows and columns of data matrix Z is asymmetrical - columns are correlated with equal weight attached to each row (station); rows are compared by distances standardized with respect to the different variables (months).

2. THE GEOMETRY AND DISPLAY OF A BATCH OF MULTIVARIATE DATA

2.1. The configuration of variables

We begin by considering the configuration of a batch, in terms of its means, standard deviations and correlations. We take it that the reader knows how to interpret each one of these measures, but that he may be bewildered by the magnitude of a correlation (or variance) matrix and may need guidance to make any sense of the, say, $\binom{20}{2} = 190$ correlations from a 20-variate data batch. We will, therefore, provide a method of representing such a configuration and show an example of interpreting it. For brevity, we will illustrate this on six-variate data.

Geometry is most useful in grasping the structure and patterns of multivariate data. One may think of the m variables as m vectors emanating from one center -- the centroid of the data -- such that (i) the length of each vector is proportional to the standard deviation of the corresponding variable, and (ii) the cosine of the angle between any two vectors is the correlation between the corresponding two variables. In fact, it follows from (1.4) and (1.5) that

$$\| \underset{\sim}{Y}_{(v)} \| = \sqrt{n}\ s_v \quad , \tag{2.1}$$

and from (1.6) that

$$\cos\ (\underset{\sim}{Y}_{(v)}, \underset{\sim}{Y}_{(v')}) = r_{v,v'} \quad . \tag{2.2}$$

Long vectors thus correspond to high variability, short ones to low variability; tight sheaves of vectors correspond to highly correlated variables; vectors at right angles to one another to uncorrelated variables; and vectors in opposite directions to negatively correlated variables.

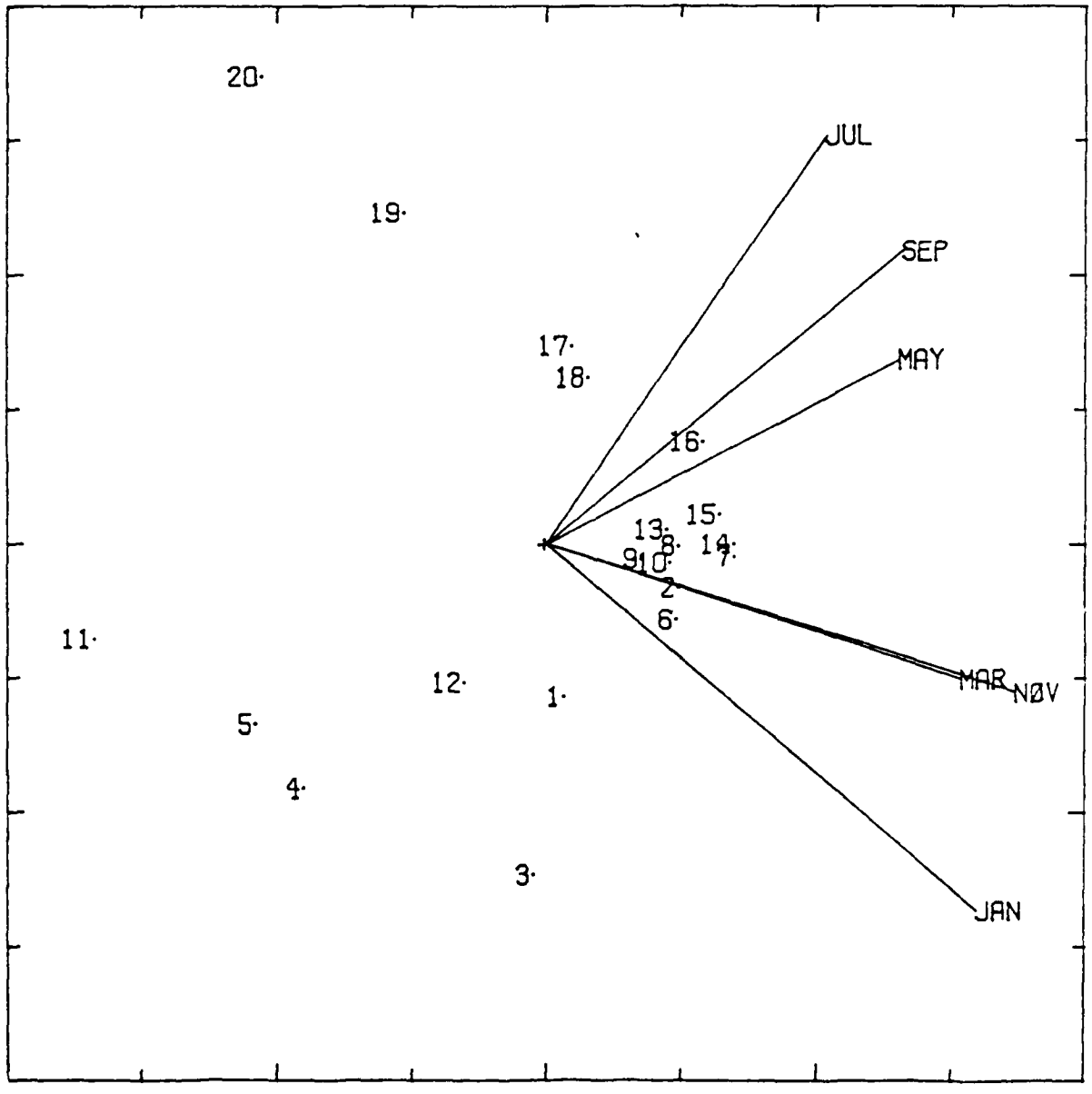
2.2. Approximation of the variables' configuration in the plane

Geometric conceptualization in hyper-space may not be to everyone's taste, but an approximate representation in the plane, or in 3D, is often quite useful in revealing many of the features of a configuration. To illustrate, consider the variances of Table 3 and the approximate representation of their configuration by the arrows of Figure 2 -- This display is called a biplot. The method of approximation will be discussed later -- subsection 2.3, below. Suffice it to say now that the goodness of fit of this planar display is 96.7% for the temperature illustration, so that little of interest could have been lost by reducing this configuration to the plane. (The dots on Figure 2 represent the stations -- more about that later -- subsection 2.6)

((Figure 2 about here))

The configuration of the arrows in the biplot of Figure 2 is particularly simple. The length of the arrows are pretty similar, indicating similar variabilities of all months; but March and May arrows are the shortest, since the standard deviations on these months are least -- see Table 1. All arrows are within the quadrant formed by those for January and July. The angle separating the latter two arrows is close to 90°: this indicates virtually zero correlation between these two months (Table 4 shows this correlation to be -.09, which is indeed negligible). In between these two

Figure 2: Biplot of Temperature Data (Table 1)



are all other months, with smaller angles, indicating positive correlation. One may describe the entire configuration roughly by two sheaves of arrows: a Fall-Winter sheaf of arrows separated by small angles, i.e., highly correlated (with March and November being particularly highly correlated), and a Spring-Summer sheaf with slightly greater angles, i.e., less highly correlated. This despription will be noted to accord completely with that obtained from Table 4, above.

The practical usefulness of the biplot is more evident when the number of variables is larger. In that case it is much easier to see patterns and sheaves on the biplot than by inspection of the matrix of correlations. It is a matter of not seeing the wood (configuration) for the trees (correlations) because there are so many of the latter. An important function of multivariate data analysis is to provide such simple descriptive tools to allow the investigator to make sense out of the mass of correlations and other data spewed out by modern computers.

2.3. Computation of planar approximations

The method of obtaining the planar approximation of the variables' configuration is to solve

$$Y'Y\underline{q}_\alpha = \lambda_\alpha^2 \underline{q}_\alpha \tag{2.3}$$

for the largest two eigenvalues $\lambda_1^2 \geq \lambda_2^2$ and their associated eigenvectors $\underline{q}_1, \underline{q}_2$ (normalized to length one). One then forms matrix

$$H_{(mx2)} = (\lambda_1 \underline{q}_1, \lambda_2 \underline{q}_2), \tag{2.4}$$

whose rows $\underline{h}_1', \ldots, \underline{h}_m'$ are plotted as arrows emanating from a common origin. This method is equivalent to least squares fitting and its goodness of fit can be gauged by coefficient

$$\begin{aligned} \lambda_{[2]}^{(4)} &= (\lambda_1^4 + \lambda_2^4)/\mathrm{tr}(Y'Y\,Y'Y) \\ &= 1 - ||Y'Y - HH'||^2/||Y'Y||^2. \end{aligned} \tag{2.5}$$

It provides the approximations

$$||\underline{h}_v|| \underline{\underline{\text{apx}}} \sqrt{n}\, s_v \tag{2.6}$$

and

$$\cos(\underline{h}_v, \underline{h}_{v'}) \underline{\underline{\text{apx}}} r_{v,v'} \tag{2.7}$$

corresponding to (2.1) and (2.2), above.

2.4. Approximation of the units' scatter in the plane

The foregoing calculations are equivalent to those of the first two principal components, a fact which will be commented on later, and they also lead to a useful representation of the units in terms of their statistical scatter. One forms

$$F_{(nx2)} = (\lambda_1^{-1}\underline{q}_1, \lambda_2^{-1}\underline{q}_2) \tag{2.8}$$

and computes matrix

$$G_{(nx2)} = Y\ F \quad , \tag{2.9}$$

whose rows $\underline{g}_1', \ldots, \underline{g}_n'$ are plotted as points. The distances between the plotted points then provide an approximate representation of the standardized statistical distances between the corresponding units, that is,

$$||\underline{g}_i - \underline{g}_i'|| \overset{apx}{=} \sqrt{d_{i,e}'}/\sqrt{n} \tag{2.10}$$

as well as

$$||\underline{g}_i|| \overset{apx}{=} \sqrt{u_{i,i}}/\sqrt{n}. \tag{2.11}$$

The coefficients obtained by performing these calculations on the temperature data are shown in Table 8.

The interpretation of such a g-scatter is obvious. Distant points represent units which are statistically dissimilar; points close together represent statistically similar units; clusters of points represent groups of

TABLE 8

Biplot (and Bimodel) Coordinates For Temperature Data

$\underline{h}_1'$	198.05	-169.37	-29.23	$\underline{f}_1'$	.0010	-.0017	-.0056
$\underline{h}_2'$	190.76	-62.32	-16.19	$\underline{f}_2'$	.0010	-.0006	-.0031
$\underline{h}_3'$	162.02	85.17	-25.32	$\underline{f}_3'$	.0008	.0009	-.0049
$\underline{h}_4'$	128.44	188.45	-13.20	$\underline{f}_4'$	.0007	.0019	-.0025
$\underline{h}_5'$	163.86	135.40	17.81	$\underline{f}_5'$	.0009	.0014	.0034
$\underline{h}_6'$	215.63	-68.44	54.51	$\underline{f}_6'$	.0011	-.0007	.0104

$\underline{g}_1'$	.0206	-.1865	-.1004	$\lambda_1 = 437.841$	$\lambda_1^2 = 191704$
$\underline{g}_2'$	.1605	-.0533	.1526	$\lambda_2 = 313.613$	$\lambda_2^2 = 98353$
$\underline{g}_3'$	-.0167	-.4055	-.3292	$\lambda_3 = 72.247$	$\lambda_3^2 = 5220$
$\underline{g}_4'$	-.2971	-.2990	.3781		
$\underline{g}_5'$	-.3549	-.2212	.0300		
$\underline{g}_6'$	.1572	-.0923	-.1148		
$\underline{g}_7'$	.2279	-.0155	-.0235		
$\underline{g}_8'$	.1604	-.0025	.1362		
$\underline{g}_9'$	.1143	-.0180	.0810		
$\underline{g}_{10}'$	.1498	-.0227	.0190		
$\underline{g}_{11}'$	-.5509	-.1157	.1811		
$\underline{g}_{12}'$	-.1000	-.1692	-.5192		
$\underline{g}_{13}'$	.1459	.0179	.1099		
$\underline{g}_{14}'$	.2273	-.0004	.0124		
$\underline{g}_{15}'$	.2092	.0362	-.1138		
$\underline{g}_{16}'$	.1897	.1254	.0817		
$\underline{g}_{17}'$	.0294	.2422	-.0913		
$\underline{g}_{18}'$	.0504	.2036	.5040		
$\underline{g}_{19}'$	-.1752	.4054	-.2196		
$\underline{g}_{20}'$	-.3480	.5711	-.1744		

Table 9

Inter-station Biplot Distances

Station	Station																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		0.86	0.99	1.51	1.69	0.74	1.20	1.03	0.86	0.93	2.58	0.54	1.07	1.24	1.31	1.59	1.92	1.75	2.79	3.77
2	0.86		1.76	2.32	2.42	0.18	0.35	0.23	0.26	0.14	3.19	1.28	0.32	0.38	0.46	0.81	1.45	1.25	2.54	3.60
3	0.99	1.76		1.34	1.72	1.60	2.06	1.97	1.83	1.87	2.72	1.12	2.03	2.11	2.22	2.55	2.90	2.74	3.70	4.61
4	1.51	2.32	1.34		0.43	2.23	2.67	2.44	2.23	2.35	1.40	1.06	2.44	2.70	2.72	2.89	2.83	2.73	3.20	3.90
5	1.69	2.42	1.72	0.43		2.36	2.76	2.50	2.29	2.43	1.00	1.16	2.48	2.78	2.77	2.89	2.69	2.63	2.92	3.54
6	0.74	0.18	1.60	2.23	2.36		0.47	0.40	0.38	0.31	3.17	1.20	0.50	0.52	0.62	0.98	1.60	1.41	2.68	3.73
7	1.20	0.35	2.06	2.67	2.76	0.47		0.31	0.51	0.35	3.51	1.62	0.40	0.07	0.25	0.65	1.45	1.26	2.61	3.68
8	1.03	0.23	1.97	2.44	2.50	0.40	0.31		0.22	0.10	3.22	1.38	0.11	0.30	0.28	0.59	1.24	1.04	2.36	3.43
9	0.86	0.26	1.83	2.23	2.29	0.38	0.51	0.22		0.16	3.01	1.17	0.21	0.51	0.49	0.72	1.22	1.03	2.09	3.35
10	0.93	0.14	1.87	2.35	2.43	0.31	0.35	0.10	0.16		3.16	1.30	0.18	0.36	0.37	0.69	1.30	1.11	2.40	3.67
11	2.58	3.19	2.72	1.40	1.00	3.17	3.51	3.22	3.01	3.16		2.03	3.17	3.52	3.47	3.48	3.05	3.04	2.87	3.20
12	0.54	1.28	1.12	1.06	1.16	1.20	1.62	1.38	1.17	1.30	2.03		1.38	1.65	1.66	1.85	1.93	1.80	2.59	3.49
13	1.07	0.32	2.03	2.44	2.48	0.50	0.40	0.11	0.21	0.18	3.17	1.38		0.37	0.30	0.52	1.13	0.93	2.25	3.32
14	1.24	0.38	2.11	2.70	2.78	0.52	0.07	0.30	0.51	0.36	3.52	1.65	0.37		0.18	0.59	1.40	1.21	2.56	3.63
15	1.31	0.46	2.22	2.72	2.77	0.62	0.25	0.28	0.49	0.37	3.47	1.66	0.30	0.18		0.41	1.22	1.03	2.38	3.45
16	1.59	0.81	2.55	2.89	2.89	0.98	0.65	0.59	0.72	0.69	3.48	1.85	0.52	0.59	0.41		0.89	0.71	2.06	3.12
17	1.92	1.45	2.90	2.83	2.69	1.60	1.45	1.24	1.22	1.30	3.05	1.93	1.13	1.40	1.22	0.89		0.20	1.17	2.24
18	1.75	1.25	2.74	2.73	2.63	1.41	1.26	1.04	1.03	1.11	3.04	1.80	0.93	1.21	1.03	0.71	0.20		1.35	2.42
19	2.79	2.54	3.70	3.20	2.92	2.68	2.61	2.36	2.29	2.40	2.87	2.59	2.25	2.56	2.38	2.06	1.17	1.35		1.07
20	3.77	3.60	4.61	3.90	3.54	3.73	3.68	3.43	3.35	3.47	3.20	3.49	3.32	3.63	3.45	3.12	2.24	2.42	1.07	

statistically similar units; sets of points ordered across the plot represent units differing in a systematic sequence, etc.

Each station i is represented by its appropriate $\underline{g}_i$ vector on the biplot of Figure 2. Note that different scales can be used for the $\underline{g}$'s and the $\underline{h}$'s -- though for each of them its horizontal and vertical scales must be the same.

Figure 2 shows that the scatter of $\underline{g}$ points mimics to some extent the geographical spread of the stations -- see map in Figure 1. Thus, the Northernmost stations appear on top of the biplot with a clear diagonal trend associated with latitude. Stations on or near the Northern coast of South America form a tight cluster (high degree of statistical similarity) whilst stations farther south and west in South America trail out towards the lower left of the biplot. If we split the stations into four geographically contiguous groups, we should expect greater homogeneity of temperature profiles within each group and considerable inter-group differences. Table 10 groups the inter-station distances of Table 7 accordingly and this confirms that the biplot clustering does produce relatively homogeneous groups.

Clearly, geographical proximity is associated with similarity in annual temperature profiles. But this association is not perfect, as witness to the fact that the west coast stations 11 and 12 are more similar to southern stations 1,3,4,5 than to stations 9 and 15 which are closer by.

TABLE 10

Median Distances Within and Between Groups

From Group	To Group				Stations
	I	II	III	IV	
I	3.3	3.6	4.1	4.3	1,3,4,5,11,12
II	3.6	2.4	3.0	3.8	2,6,7,8,9,10,13,14,15
III	4.1	3.0	2.0	1.9	16,17,18
IV	4.3	3.8	1.9	2.8	19,20

2.5. Plotting of extra data points

The biplot has been constructed to display a data matrix Z about its centroid $\bar{z}'$, that is, it represents $\bar{z}'$ at its origin and displays deviations Y. At times one may wish to display further data points that were not fitted in obtaining the biplot. Thus, one may have an additional unit with m-variate observations z_0'. This will be centered as deviation $y_0' = z_0' - \bar{z}'$ and its biplot coordinates calculated as

$$g_0' = y_0' F \tag{2.12a}$$

or

$$g_0' = (z_0' - \bar{z}')F \tag{2.12b}$$

To illustrate, one might consider a hypothetical station whose temperatures for each month were exactly one standard deviation above the mean, i.e.,

z_0' = (284.68, 274.58, 273.46, 281.42, 287.14, 280.65).

Calculation of (2.12) yields biplot coordinates g_0' = (.2748, .0379). Such a point would appear in the biplot -- Figure 2 -- slightly to the right of g_{15}. Indeed, the temperatures for station 15 are similar but slightly smaller than those of this hypothetical station.

2.6. Joint variables and units approximation -- the biplot

The biplot (Gabriel, 1971) displays both the configuration of the months (variables - columns of data matrix Z) and the scatter of stations (units - rows of Z). Because it displays them jointly it is called a biplot -- and this simultaneous representation allows more insight into the data than could be obtained from the separate inspections of variables (subsection 2.3) and of rows (subsection 2.4) which have been illustrated above.

The biplot displays the actual deviations $y_{i,v} = z_{i,v} - \bar{z}_{(v)}$ by inner products

$$y_{i,v} \underline{\underline{\text{apx}}} \; \underline{g}_i' \underline{h}_v \quad , \tag{2.13}$$

with goodness of fit

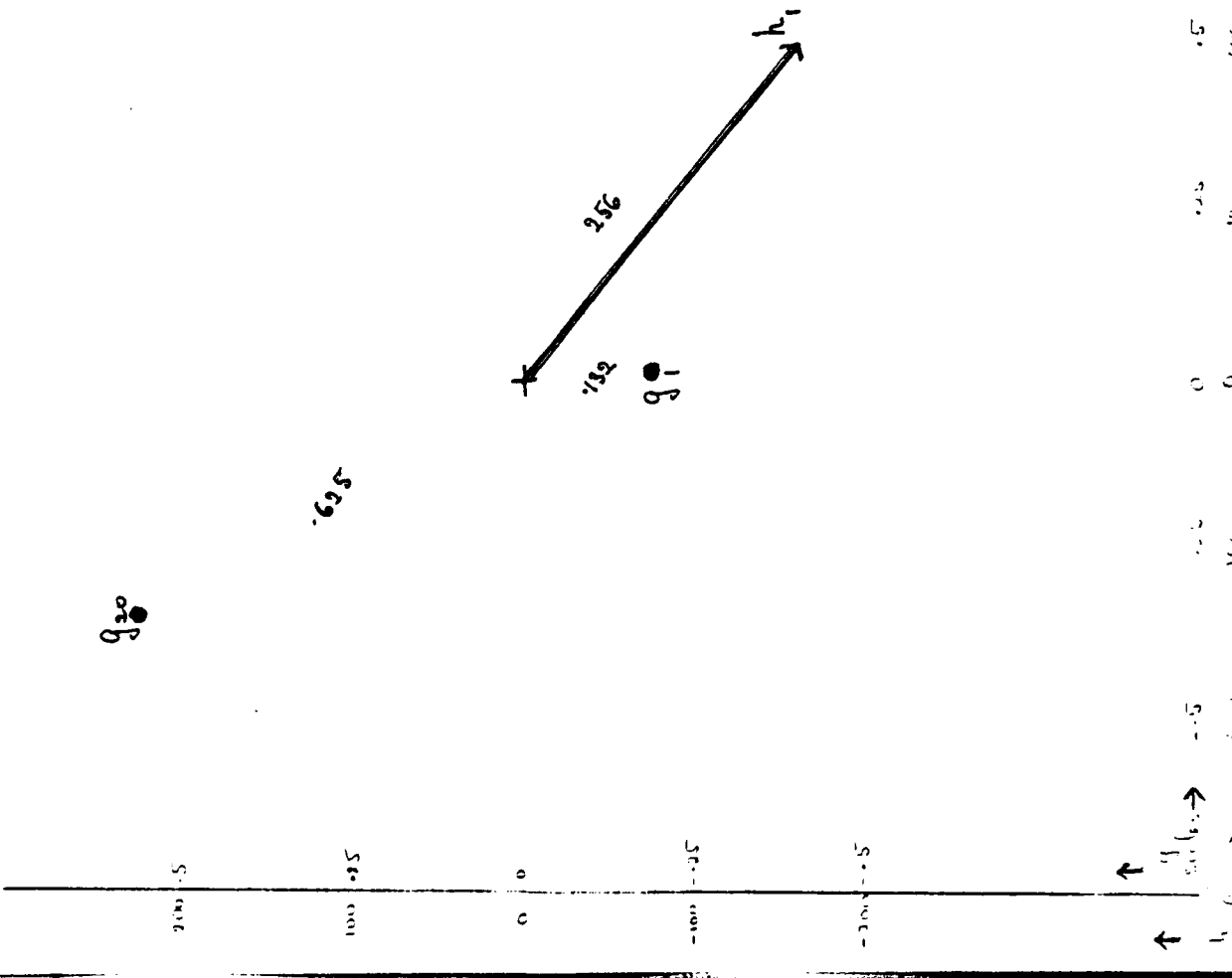
$$\begin{aligned} \lambda^{(2)}_{[2]} &= (\lambda_1^2+\lambda_2^2)/\text{tr}(Y'Y) \\ &= 1 - ||Y-GH'||^2/||Y||^2 \quad . \end{aligned} \tag{2.14}$$

In other words, the deviation for station i on variable v can be visualized as the length of $\underline{h}_v$ times the length of the projection of $\underline{g}_i$ (considered as a vector from the origin) onto $\underline{h}_v$ -- the sign of the lengths' product being positive or negative according to whether $\underline{g}_i$'s projection onto $\underline{h}_v$ is in the same or opposite direction to $\underline{h}_v$ itself. Clearly, then, a station i whose $\underline{g}_i$ is far out in the direction of (opposite to) the vector $\underline{h}_v$ of a variable v has large positive (negative) deviation $y_{i,v}$. When the $\underline{g}_i$'s are less far out in the $\underline{h}_v$ direction (or opposite it) the deviations are smaller.

Figure 3 displays the January arrow $\underline{h}_1$ and the station 1 and station 20 points $\underline{g}_1$ and $\underline{g}_{20}$ of the biplot of Figure 2. It also shows the orthogonal projection of these two points onto the line through $\underline{h}_1$. The projection of $\underline{g}_1$ is seen to be of length 0.132 in the direction of $\underline{h}_1$ -- whose length is 256. Hence the biplot approximation of $y_{1,1} = 34.4$ is $\underline{g}_1'\underline{h}_1 = .132 \times 256 = 33.8$. Similarly, the projection of $\underline{g}_{20}$ onto the line through $\underline{h}_1$ is of length 0.625 in the direction opposite $\underline{h}_1$. Hence the biplot approximation of $y_{20,1} = -157.6$ is $\underline{g}_{20}'\underline{h}_1 = -.625 \times 256 = -160$ (the minus sign being attached because the projection is opposite the vector projected upon).

This relation between $\underline{g}_i$ points and $\underline{h}_v$ arrows is useful in interpreting the scatter of $\underline{g}$ points. Thus, one may identify the variables (months) on which a cluster of units (stations) is particularly large or small. As an example, we note the northernmost stations in Figure 2 to be aligned in a direction opposite the Fall-Winter sheaf. Evidently, the farther north the station, the lower its Fall-Winter temperatures. (This is readily confirmed by inspecting the last four or five rows of Table 2.) On the other hand, the difference between the second and third clusters of stations is associated with the direction of the Spring-Summer temperatures: the nortn coast stations have higher Spring-Summer temperatures than the western and southern South American stations. (Again, Table 2 confirms this pattern.)

Figure 3: Biplot Reconstruction of Two Observations

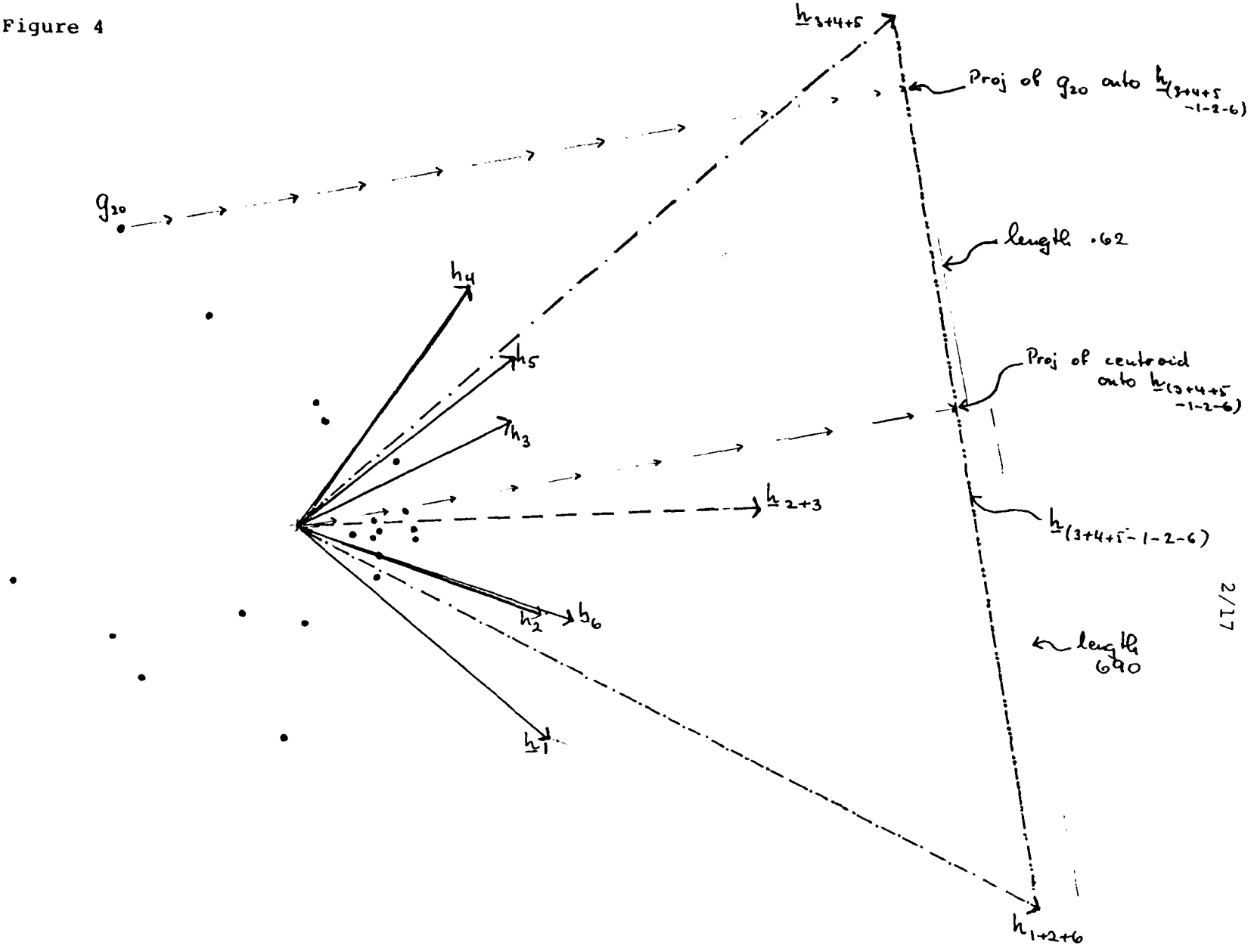


At this stage it is well to realize that one can inspect the biplot also for linear combinations of variables beyond the actual variables of the data displayed. One may do this by vector addition of $\underline{h}$ arrows. Thus, for example, a March plus May sum would be represented by the vector $\underline{h}_2 + \underline{h}_3$ which is readily constructed on the biplot -- the dashed line on Figure 4 -- and found to be roughly horizontal. Also, a Spring-Summer sum ($\underline{h}_3 + \underline{h}_4 + \underline{h}_5$) -- dashed and dotted line on Figure 4 -- slants up at roughly 45° whereas a Spring-Summer versus Fall-Winter difference ($\underline{h}_3 + \underline{h}_4 + \underline{h}_5$) - ($\underline{h}_1 + \underline{h}_2 + \underline{h}_6$) -- dashed and double dotted line on Figure 4 -- is pretty close to vertical.

The importance of such combinations of variables is great. For example, we note the northern hemisphere station points to be mostly above the biplot origin and the southern hemisphere station points to be below. This vertical difference evidently is one of Spring-Summer versus Fall-Winter temperatures -- the very well known fact that maximum temperatures in the Northern(Southern) hemisphere are in the Spring-Summer(Fall-Winter). Similarly, the north coastal South American station points are farthest to the right of the biplot, indicating that average temperatures are highest in that region -- again a well-known fact.

These features of the biplot are of considerable importance for data analysis. They allow one to go beyond

Figure 4


h_{3+4+5}
Proj of g_{20} onto $h_{(3+4+5 -1-2-6)}$
g_{20}
length .62
h_4
h_5
Proj of centroid onto $h_{(3+4+5 -1-2-6)}$
h_3
h_{2+3}
$h_{(3+4+5-1-2-6)}$
h_2
h_6
length 690
h_1
h_{1+2+6}

2/17

separate descriptions of variables and of units and actually account for units' clusters and patterns in terms of the variables that determine them.

Finally, it will be noted that the signed length of the projection of any unit's $\underline{g}_i$ vector onto any variable's $\underline{h}_v$ vector direction approximates $1/\sqrt{n}$ times the standardized (mean zero, variance one) observation on that variable, i.e.,

$$\underline{g}_i'\underline{h}_v/||\underline{h}_v|| \underline{\underline{apx}} (z_{i,v} - \bar{z}_{(v)})/\sqrt{n}\, s_v \qquad (2.15)$$

--this is evident from approximations (2.6) and (2.13) of s_v and $y_{i,v}$, respectively. Clearly, the same holds for any linear combination of variables when projections are made onto the appropriate vector combination of $\underline{h}$'s.

To illustrate, the centroid to $\underline{g}_{20}$ vector has been projected onto $\underline{h}_{(3+4+5-1-2-6)}$ in Figure 4. The length of the projection is .62 whereas the length of the vector projected upon is 690. Thus the biplot approximation of the standardized Spring-Summer versus Fall-Winter difference for station 20 is $\sqrt{20}$ x .62 = 2.77 -- this is an extreme observation as is evident from the biplot and the actual difference 463.8 (as measured from the centroid) is approximated on the biplot by .62 x 690 = 428.

2.7. Joint approximation in three dimensions - the bimodel

The biplot displays the rank 2 least squares approximation of Y by GH'. One could similarly obtain a rank 3 approximation by solving (2.3) and (2.8) also for $\alpha=3$ and adding a further column to H, to F and to G -- the resulting bimodel could be constructed in three-space since each $\underline{g}_i$ and $\underline{h}_v$ now has three coordinates. Higher dimensional approximations can also be calculated by solving (2.3) and (2.8) for further α's, but these cannot be constructed physically.

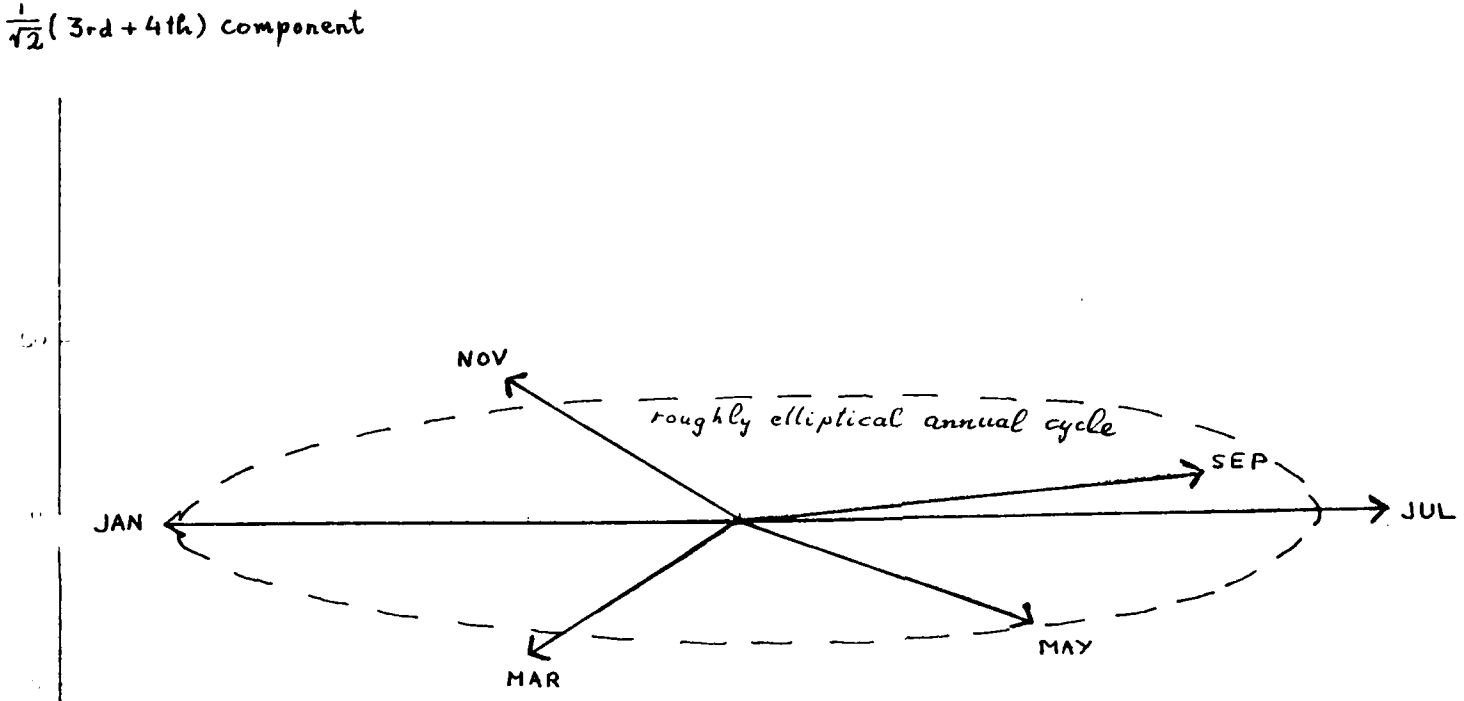
It is, however, feasible to inspect the three or higher dimensional approximations by displaying various projections on a CRT. Facilities exist on some computer installations that allow rotation of the higher dimensional approximation so one gets successive two dimensional views from different angles. This may be quite useful in revealing features of data that are not apparent from the original planar approximation.

As an illustration, consider again the $\underline{h}$ configuration in Figure 2. The annual cycle is represented by an upward movement from $\underline{h}_1$ through $\underline{h}_2$ and $\underline{h}_3$ to $\underline{h}_4$ and then a pretty similar downward movement from $\underline{h}_4$ through $\underline{h}_5$ and $\underline{h}_6$ to $\underline{h}_1$. This suggested that there might be something like an elliptical orbit of the $\underline{h}$'s in three-space. Indeed, if one replots the $\underline{h}$ arrows along the second principal axis and an axis at 45° to the third and fourth principal axes (found by trial and error) are does find such an orbit --

Figure 5. True, the extra axis displayed in Figure 5 accounts for very little of the data's variability, but there is something satisfying to have a model which displays an annual cycle rather than a mere two season clustering.

Whether or not such a model is appropriate and worthwhile for the data used here, it illustrates the possibilities of using the higher dimensional bimodels for further inspection of data.

Figure 5: h arrow configuration of temperatures along special axes



3. DATA ANALYSIS OF THE VARIABLES' CONFIGURATION

3.1. Purposes of data analysis

Data analysis aims at systematizing and summarizing data by noting regularities, tracing patterns, fitting models, etc. When the configuration of a set of variables is described by its variance matrix, a data analysis will attempt to elicit the salient features of variability and inter-correlation of the variables. Display of $\underline{h}$-vectors in a biplot, or in a higher dimensional bimodel, allows visual inspection of the configuration and may suggest grouping of highly correlated subsets of variables whose $\underline{h}$-vectors form tight sheaves. It may also indicate regular patterns such as the elliptical orbit associated with the annual cycle of temperatures illustrated above (Section 2.7). Such indications of regularity, whether suggested by visual inspection or otherwise, may lead to formulation of a "model" or systematic description of the set of variables.

3.2. Variables' sheaves and clustering algorithms

The most common concept used for such descriptions is that of a "typical variable." When the variables group naturally into subsets such that there is high correlation between variables within subsets and much lower correlation from subset to subset, then one naturally thinks of a "typical" variable for each subset. Geometrically, when the h-vectors separate into several tight sheaves, one may well describe each sheaf by a typical, or average, h-arrow going through the center of the sheaf. Thus, in the temperature example in Section 2.2 above, the months' configuration seemed to cluster into a Fall-Winter sheaf and a Spring-Summer sheaf, and one could think of a typical variable for each.

Where there are too many variables for easy direct or graphical inspection, one may try to check for sheaves by some method of cluster analysis. If the variables do form separate tight sheaves, this will be revealed by any clustering algorithm. However, in many cases there are no tight and well separated sheaves and application of clustering techniques does not yield meaningful results.

Unfortunately, clustering algorithms - of which there are many, and quite a few are available within standard statistical computer packages - always produce some output. When no clear sheaves exist in the configuration, different algorithms will yield different clusterings, none of which

are particularly meaningful. In such cases one would do better not to use algorithms to force "clustering" -- they should be used only to check if clustering actually exists and then reveal the existing clusters. A good practical rule may be to use a number of alternative "clustering algorithms": any "clusters" that are not revealed by most of the algorithms must be considered suspect - they are likely to be artifacts.

3.3. Principal components

When variables do not readily group into distinct sheaves one may define "typical" variables, or LCV's, in a different sense, one more akin to averaging. Thus, the "most typical" LCV is often taken to be that which has the highest average correlations, or squared correlations, with the observed variables. This is obviously an attractive descriptive property - such a "most typical" LCV is by definition highly correlated with the variables it typifies.

If one wishes to describe the variables' configuration by more than one "typical" LCV one may consider the residuals *from regressing* each variable on the first "most typical" LCV. Again, one may seek the LCV most highly correlated with the residual parts of the variables - this will be the "second most typical" LCV and will be found to be uncorrelated with the first.

One may continue in this way, again taking residuals and obtaining a "third most typical LCV," etc.

The logic of looking at these successive residuals is not straightforward. Only the first of the "typical" LCV's is directly related to the original variables. For all the others it is not at all obvious if they can be considered "typical" of the original variables.

When the criterion of "most typical" is that of maximum average squared covariance for any normalized LCV (i.e., an

LCV $\underline{Y}_{(\underline{c})} = \Sigma_v c_v \underline{Y}_{(v)}$ normalized so that $\Sigma_v c_v^2 = 1$) the resulting "typical variables" are referred to as <u>principal components</u>. Thus, the first principal component (PC_1 for short) is the vector

$$\Sigma_v c_v Y_{(v)} = Y\underline{c} \tag{3.1}$$

which satisfies

$$\max_{\underline{c}} \{\Sigma_v (\underline{Y}_{(v)}' Y\underline{c})^2 : \underline{c}'\underline{c} = 1\}$$

$$= \max_{\underline{c}} \{||Y'Y\underline{c}||: \underline{c}'\underline{c} = 1\}. \tag{3.2}$$

But this is satisfied by solution $\underline{c} = \underline{q}_1$ of equations (2.3). Hence PC_1 is given as

$$Y\underline{q}_1 = \lambda_1 \underline{p}_1 \quad , \tag{3.3}$$

which is the first column of G (2.8),

The next solution of (2.3), i.e., $\underline{q}_2$, similarly yields PC_2 as

$$Y\underline{q}_2 = \lambda_2 \underline{p}_2 \quad , \tag{3.4}$$

the second column of G.

These two PC's are uncorrelated for

$$\lambda_1 \underline{p}_1' \underline{p}_2 \lambda_2 = \underline{q}_1' Y' Y \underline{q}_2 = \underline{q}_1' \underline{q}_2 \quad . \tag{3.5}$$

But these eigenvectors are known to be orthogonal unless $\lambda_1 = \lambda_2$.

In general, PC_α is the LCV with observations vector $\lambda_\alpha \underline{p}_\alpha$ obtained by the solution of (2.3) with the α-th largest root λ_α^2.

Another property of the PC's is that PC_1 has the largest variance of all normalized LCV's, i.e.,

$$||Y\underline{q}_1||^2/n = \lambda_1 \underline{p}_1' \underline{p}_1 \lambda_1/n = \lambda_1^2/n \quad . \tag{3.6}$$

Similarly, amongst all LCV's uncorrelated with PC_1, it is PC_2 which has the largest variance

$$||Y\underline{q}_2||^2/n = \lambda_2 \underline{p}_2' \underline{p}_2 \lambda_2/n = \lambda_2^2/n \quad , \tag{3.7}$$

and so on for other PC's. Another property of PC's is that the first two provide the principal axes of the biplot - PC_1 is along the horizontal axis, PC_2 along the vertical - and the remaining PC's are along the other principal axes of the bimodel and higher order approximations. This last property is due to the fact that PC's have simple least squares properties (which were discovered by Householder and Young in 1938). In particular, the plane that best approximates the configuration is that going through the first two principal axes. In other words, the best fitting two dimensional approximation of Y is

$$Y_{[2]} = \lambda_1 \underline{p}_1 \underline{q}_1' + \lambda_2 \underline{p}_2 \underline{q}_2' \quad , \tag{3.8}$$

which is a function of PC_1 and PC_2 only. It is because of this least squares property that these two principal axes were chosen to serve as horizontal and vertical axes of the biplot.

The generalization of these remarks to a 3-D or higher order approximations is obvious.

The relation of the PC's to coordinates of the biplot is simply that the i-th unit's observation on PC_α is

$$\underline{Y}'_i \underline{q}_\alpha = \lambda_\alpha g_{i,\alpha} \quad . \qquad (3.9)$$

For the first two PC's this follows from (2.9) and (3.3), (3.4). Table 11 gives the six principal coordinates for the twenty stations of the temperature data of Table 1. It is not obvious what interpretation one might want to put on these coordinates as such, though plotting the first three -- scaled by λ_1, λ_2 and λ_3, respectively, was found useful in inspection of the bimodel.

Table 11

Principal Component Scores

	$\lambda_1 p_{1,i}$	$\lambda_2 p_{2,i}$	$\lambda_3 p_{3,i}$	$\lambda_4 p_{4,i}$	$\lambda_5 p_{5,i}$	$\lambda_6 p_{6,i}$
1	9.0	-58.5	-7.3	-7.5	-0.2	0.2
2	70.3	-16.7	11.0	-10.6	-19.0	-2.0
3	-7.3	-127.2	-23.8	41.5	-7.2	1.7
4	-130.1	-93.8	27.3	-4.4	-0.7	-1.0
5	-155.4	-69.4	2.2	2.6	8.5	1.4
6	68.8	-29.0	-8.3	-11.0	3.5	-2.0
7	99.8	-4.9	-1.7	-1.4	2.1	-7.3
8	70.2	-0.8	9.8	2.4	1.0	-0.9
9	50.0	-5.7	5.8	-0.6	3.7	-0.4
10	65.6	-7.1	1.4	-2.1	-10.4	7.3
11	-241.2	-36.3	13.1	-12.3	-0.8	0.9
12	-43.8	-53.1	-37.5	-14.0	5.0	-3.2
13	63.9	5.6	7.9	-1.7	2.0	-2.5
14	99.5	-0.1	0.9	-17.0	-1.7	5.7
15	91.6	11.3	-8.2	0.5	9.7	-1.9
16	83.1	39.3	5.9	7.6	11.4	3.5
17	12.9	76.0	-6.6	-1.4	1.5	7.7
18	22.08	63.9	36.4	21.4	2.3	-3.1
19	-76.7	127.1	-15.9	2.3	-12.8	-5.1
20	-152.3	179.1	-12.6	5.9	2.1	1.0

3.4. Principal component analysis

None of these mathematical properties make it clear why the PCs should be particularly interesting for an understanding of the variables' configuration. Clearly, PC_1 makes intuitive sense as an "average" or typical variable, or as the LCV with maximal variability. But what of PC_2, PC_3, etc.? They do not seem to have clear intuitive descriptive appeal. Their least squares property makes them useful for building approximations in the plane - the biplot -, in 3-D - the bimodel -, etc., but that does not make them interesting individually. In the temperature example, the interesting "typical" LCVs seemed to be going at 45° and -45° rather than at 0° and 90° to the horizontal. The fact that the PC's and principal axes are useful for plotting does not necessarily make them useful for interpretation. One usually does better by relating the g-points to the $\underline{h}$'s of the original variables rather than to the axes for the PC's.

A great many applications of PC analysis have been made through the years. As a method of approximation in lower dimensional space, this is fine, but as an interpretative device its popularity is surprising. What the method does is to express the original variables in terms of PCs in the form

$$\underline{Y}_{(v)} = \lambda_1 \underline{p}_1 q_{1,v} + \lambda_2 \underline{p}_2 q_{2,v} \qquad (3.10)$$

- as follows by taking a column of equation (3.8). In view of (3.3), (3.4), it is clear that the method also allows the PC's to be expressed in terms of the variables as

$$\lambda_\alpha \underline{p}_\alpha = \Sigma_v \underline{y}_{(v)} q_{\alpha,v} \quad . \qquad (3.11)$$

(Note that the weights in both linear combinations are the q's obtained by solving (2.3): they are therefore referred to as <u>loadings</u>.) But does this provide any insight? Does (3.10) "explain" the variables by the PC's or does (3.11) "explain" the PC's by the variables? Or do we expect, by circular reasoning, to have both "explanations"?

At best, consideration of the loadings q gives some insight into which variables are correlated with what others (similar loadings on the first few PCs). What is puzzling are the attempts of many users of PC's to "reify" these mathematical constructs and ascribe "inherent," "underlying" or "explanatory" content to them. It is not evident how any such content follows from the mathematical definition of PC's and one may suspect that much of what has been published as PC analyses may have obscured rather than illuminated the configuration of the original variables which should have been studied.

The loadings of the monthly temperatures in the six PC's are shown in Table 12. The uniformly high PC_1 loadings for all months, show PC_1 to be some average annual

TABLE 12

Temperature Data - PC Loadings $q_{i,\alpha}$

Variable	i	PC α 1	2	3	4	5	6
JAN	1	.452	-.540	-.405	-.561	-.158	-.005
MAR	2	.436	-.199	-.224	+.700	.028	-.479
MAY	3	.370	.272	-.350	+.258	.088	.770
JUL	4	.293	.601	-.182	-.349	.505	-.377
SEP	5	.374	.432	.247	-.073	-.772	-.107
NOV	6	.492	-.218	.754	-.034	.339	.156

temperature factor which puts more emphasis on Fall-Winter -- as is evident from the biplot in which all months' $\underline{h}$-arrows point left, partly above and partly below the horizontal direction, but the Fall-Winter $\underline{h}$'s are closer to horizontal than the Spring-Summer $\underline{h}$'s. The PC_2 loadings are positive for Spring-Summer, negative for Fall-Winter, and thus indicate a seasonal component -- again, this was evident from the biplot configuration. Loadings on PC_3, PC_4, PC_5 and PC_6 are not so easily interpretable -- though the joint consideration of PC_2, PC_3 and PC_4 in a bistructure could be interpreted and modelled in terms of an annual cycle. Nothing is revealed by consideration of these axes that was not seen by inspection of the $\underline{h}$'s themselves. The $\underline{h}$-configuration is much more simply described by two sheaves, one for each half-year, than by two axes, one a weighted annual average, the other a weighted contrast.

This example illustrates the shortcomings of PC analysis in considering each principal axis separately and the advantage of seeing the overall picture in a biplot or bimodel. It also illustrates the limitations of using orthogonal axes fitted by least squares -- these do not necessarily provide the most readily interpretable references.

3.5. Some further comments on principal components

PC's depend on the scale of measurement of the original variables. This dependence is obvious from the definitions which depend on covariances, variances and least squares fits. Much has been written on this dependence and how it limits the usefulness of PC analysis. All this seems beside the point: There is no real reason to consider PC analysis as a method for revealing the "underlying" structure or to regard PC's as "intrinsic" variables and hence there it also does not matter that these "structures" and "intrinsic variables" are not scale independent.

Finally, PC's are often reified by reference to other variables extraneous to the original set. In the temperature illustration, it was not difficult to label PC_1 and PC_2, though no simple interpretation was evident for other PC's. Another illustration is a recent study of rainfall where PC_4 was noted to have a clear time trend associated with the spread of irrigation. It was therefore suggested that PC_4 was an "irrigation factor" and the possibility of using it as such a variable was considered. Direct use of the irrigation data -- with which PC_4 had been found to be correlated -- would have been simpler and more straightforward and would not have required PC analysis at all. This was a pretty typical example of the "use" of PC's and the rationale of many other uses of PC analysis is equally puzzling.

To summarize, this author's view is that PCs are unlikely to have explanatory value in themselves. The most physically meaningful LCVs will not usually happen to lie along the principal axes of the configuration. This author sees the main usefulness of PCs as a tool to provide least squares approximations to data matrices and variables' configurations and he would direct the scientific attention of investigators to what the approximation tells them about the original variables, and not to what it shows about the PCs. The investigator must have included his variables because he wants to know something about them, so let him discuss them instead of substituting mathematical artifacts.

3.6. The rank or dimension of a configuration

When a number m of variables are observed on more than m units, the configuration may be completely in a sub-space of m-1, m-2 or fewer dimensions, but this is very unusual. It indicates exact linear relations between the variables, even though these variables must be affected by random variability and measurement error. When such things are actually observed, one usually finds that the original set of variables includes some repetitions of observations or sums, or averages, of other variables which are also included. It is rare and surprising to find such exact dependence otherwise. A set of m variables observed on n (>m) units almost invariably generates a configuration in m-space that cannot fit exactly into any lower dimensional subspace. It may well be approximated in a lower space, and perhaps even very closely approximated, but it is very unlikely to fit exactly.

This suggests that the question of dimensionality rarely relates to the true configuration of variables but usually makes sense only in the context of approximation. "Hypotheses" of reduced dimensionality are, in this author's opinion, rank nonsense and the techniques of statistical significance are only rarely relevant to problems of dimensionality. "Tests of significance" of PC's will therefore not be discussed here. If the hypothesis that a 6-variate

configuration is in a plane is physically incredible, it makes no sense to test it for significance, i.e., to test the nullity of PC_3, PC_4, etc. Hypotheses testing makes sense only if the hypotheses can be given credence.

This issue of dimensionality should correctly be addressed as one of approximation and not of hypothesis testing. The biplot plane fitted the 6-variable temperature data to a goodness of fit of 0.967 and the 3-D bimodel had a 0.985 fit. That may well justify ignoring the remaining dimensions to all intents and purposes even if one is certain that there is some variability along those axes. There is no question of "testing" whether the data are in a plane or in 3-D; the practical issue is simply that the fraction of real variation that lies outside the plane or 3-D is negligible and need not be considered in interpreting the data - even though it is not assumed to be strictly null.

3.7. The factor analytic model

Factor analysts postulate a model in which each variable can be written in the form

$$\underline{y}_{(v)} = \Sigma_{\alpha=1}^{r} \underline{f}_{(\alpha)} l_{\alpha,v} + \underline{e}_{(v)} , \qquad (3.12)$$

as the sum of a linear combination of a few, say r, "factor" variables $f_{(1)}, f_{(2)}, \ldots$ with "loadings" $l_{\alpha,v}$ and errors $\underline{e}_{(v)}$ specific to each variable and uncorrelated either with the factors or with one another.

When the rank r is sufficiently large compared to m, this model has an exact solution. In fact it then usually has infinitely many solutions. However, factor analysts usually postulate r to be quite small relative to m (so as to obtain parsimony in description), and then the model is most unlikely to fit exactly. A satisfactory fit may at times be obtained if the data are considered as a random sample from a population in which the $\underline{e}$'s are uncorrelated between themselves and with the f's.

What factor analysts do in practice is to approximate the data by a model of type (3.12) with rank as low as will allow a reasonable fit. When the purpose of a factor analysis is avowedly approximative the criteria for the method would be goodness of fit and parsimony. If we compare an approximation by model (3.12) with a PC approximation of the same rank, it is clear that the factor analytic model is very much less parsimonious in that it requires the

$\underline{e}$ terms to be uncorrelated. The fit of its $\Sigma_\alpha \underline{f}_\alpha \underline{l}_\alpha$ part is necessarily worse than that of the first r PCs because the latter are required only to give the least squares fit.

However, one variant of Factor Analysis sets out directly to approximate the correlations. Unlike PC analysis it does not approximate the data matrix Y, nor does it approximate the diagonal elements of R as these are known to equal unity. This particular variant of factor analysis -- called MINRES -- is justified directly in terms of optimal approximation of the correlations - the off-diagonal elements of R.

As in the case of principal components, one has to ask whether the model as such makes physical sense so that the factors are "intrinsic" variables, or whether it serves as a mere vehicle of parsimonious approximation. Our answer to the first question should be similar to the one we have given for PC analysis: We see no a priori reason to think the "factors" fitted in model (3.12) are any more "real" than the PCs.

The factor analytic model is no more plausible than the hypothesis of lower dimensionality which we discussed in connection with PC analysis. However, its saving grace is that it is so flexibly defined as to allow considerable manipulation which can on occasion be used to advantage.

Thus, for given rank r, model (3.12) becomes

$$Y - E = F\ L, \tag{3.13}$$

with obvious definitions for matrices E, F, and L, and the flexibility of this representation is that not only E is not uniquely determined, but F and L can be changed into

$$F^* = F\ Q \tag{3.14}$$

and

$$L^* = Q^{-1}L \tag{3.15}$$

by any non-singular rxr matrix Q. Factor analysts spend much ingenuity in rotating their original F, L solution into a solution F*, L* that "makes sense" so that the resulting $\underline{f}^*_\alpha$ "factors" have some reality and are useful in interpreting data.

In some cases these rotations are chosen so as to yield factors correlated with extraneous variables or other information available to the investigator. It is difficult to see what "explanatory" function such a procedure has. The investigator had the "explanation" or extraneous variable anyway, and he could have correlated the original variables with it. Why bother to use factor analysis? Why not just take the multiple regression of the extraneous variable on the $\underline{Y}_{(v)}$'s as the "factor"?

Some methods of rotation such as varimax and other computerized techniques are built so as to make individual $\underline{f}^*_\alpha$'s as closely representative of sheaves of variables as

possible. That brings us back to the subject of applying clustering techniques to variables' configurations, an approach whose careful use may well yield important data analytic insight.

Again, it is difficult to see what the role of the factor analytic model is in all this. If one seeks correlation with extraneous information one can best do it directly, on the data rather than on the "factor solution." If one wants to organize the variables into sheaves, it is not obvious that one had best start from a set of fitted loadings - but it may be legitimate to do so. It is essential to understand that in all these applications there seems no essential role to the "factor analytic model." This model has neither reality nor much usefulness, except under certain circumstances, as an approximating device.

Our view of factor analysis differs sharply from that of most practitioners of these techniques who talk about their model as though it had inherent reality. Even when they use a clearly approximative technique such as MINRES they try to reify the resulting factors. Indeed, the MINRES solution would sometimes involve imaginary numbers (Gabriel, 1978) but factor analysts shy away from such an optimal approximation because they believe in the "reality" of their factors.

4. ANALYZING THE SCATTER OF THE UNITS

4.1. A batch of units and its scatter

The units whose observations make up the rows of data matrix Z usually have identities of their own -- "labels" as the sampling theorists call them -- and these identities may be relevant to the analysis of the data. Some relations between units may be given a priori and it may be of interest to study if and how they are associated with statistical similarity of the corresponding rows of Z. A priori groupings of the units in terms of information, extraneous to data matrix Z, could be related to the statistical scatter and to similarities of the corresponding $\underline{z}_i$'s. Data analysis is often concerned as much with the units as with the variables. In our example it is certainly as legitimate, and interesting, to study the scatter of stations as it is to study the variance configuration of months.

In modern statistics books this subject is hardly dealt with at all, and the idea of between units distance barely receives mention. This is because the fashion has been to deal exclusively with inference based on random samples from a population or distribution. And in that context the units of observation lose their individuality and become mere replications in a sampling process.

This sampling approach is undoubtedly appropriate in many experimental situations and in situations like industrial quality control where repeat observations are carried out regularly. But it is not appropriate to the study of batches of units with well defined identities and labels. Ignoring the information associated with these identities may stultify the analysis of such data.

In this section we consider, therefore, methods of analyzing the units' scatter and we choose to do so in terms of standardized distances $\sqrt{d_{i,i'}}$ between pairs of units (1.13) and $\sqrt{u_{i,i}}$ between units and the centroid (1.12). We will find it convenient to consider the scatter also in terms of the biplot approximations $||\underline{g}_i - \underline{g}_{i'}||$ and $||\underline{g}_i||$ - (2.10) and (2.11) - of the above distances.

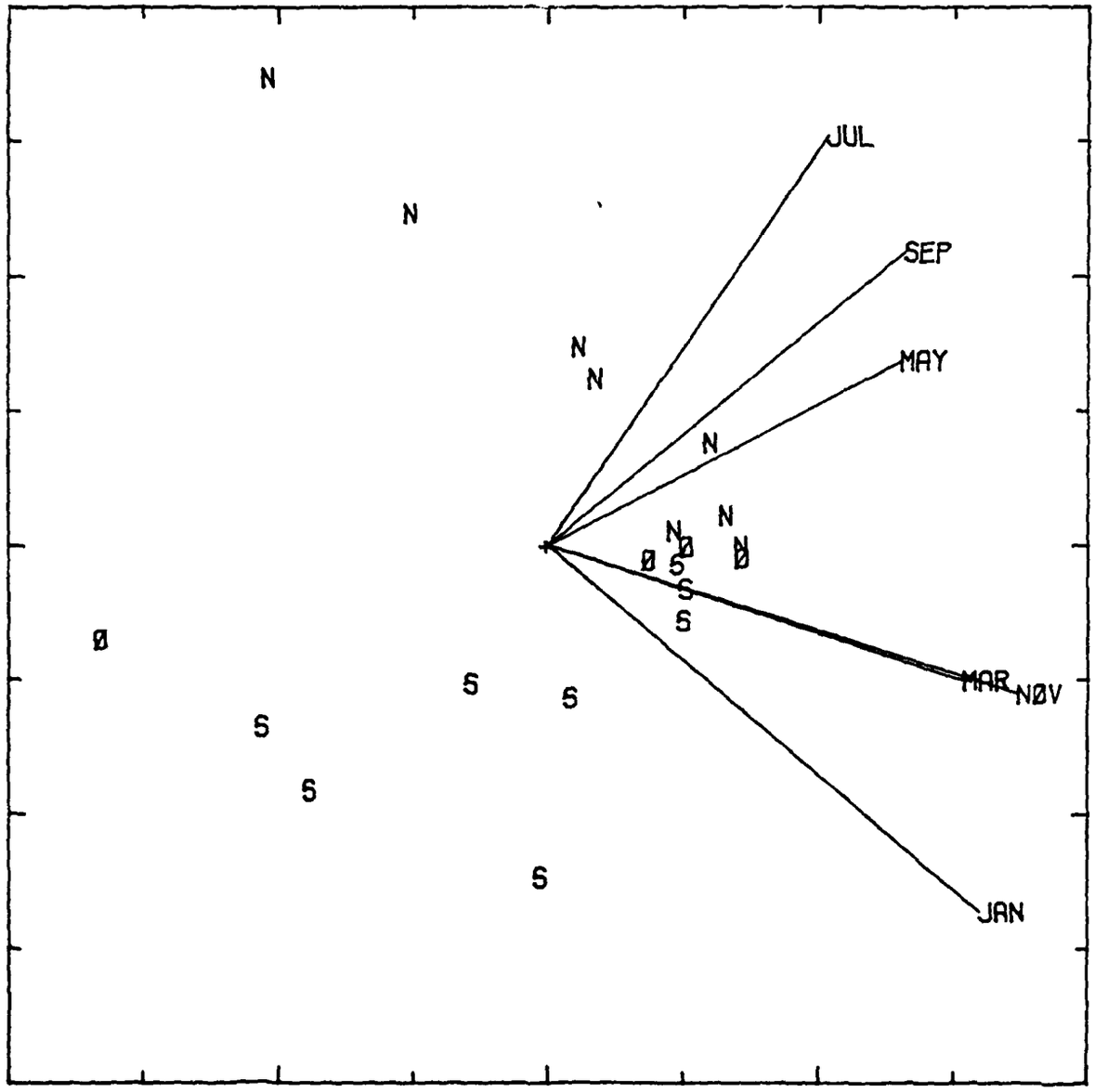
4.2. Use of extraneous information on the units

When extraneous information is available, i.e., information on unit i other than the observation $\underline{z}_i'$, this information may be correlated with the observations. When the units fall into a number of categories, one may check whether these categories are associated with the statistical scatter of points. Do the categories form distinct groupings in m-space and/or on the biplot? Is there much or little overlap between categories?

A simple device is to mark the units of each category by a different mark, or color, on the biplot and see if the categories do separate. Figure 6 shows the g-points of the temperatures biplot (Figure 2) classified according to whether they are North or South of the equator. A clear separation is evident, showing that the temperature profiles of Northern hemisphere stations differ from those in the Southern hemisphere. The former are at the top of the biplot, the latter at the bottom. Recalling that the vertical direction on the biplot was a contrast between Spring-Summer and Fall-Winter (Section 2.5, above) one sees that the Northern versus Southern hemisphere groupings reflect the difference in the season in which their maximum temperatures occur.

When the extraneous information is not categorical but rather of a continuously variable character, the methods of

Figure 6: Biplot of Temperature Data With Indication of Hemisphere of Station (N = Northern; 0= Within 2 Degrees of Equator; C = Southern)



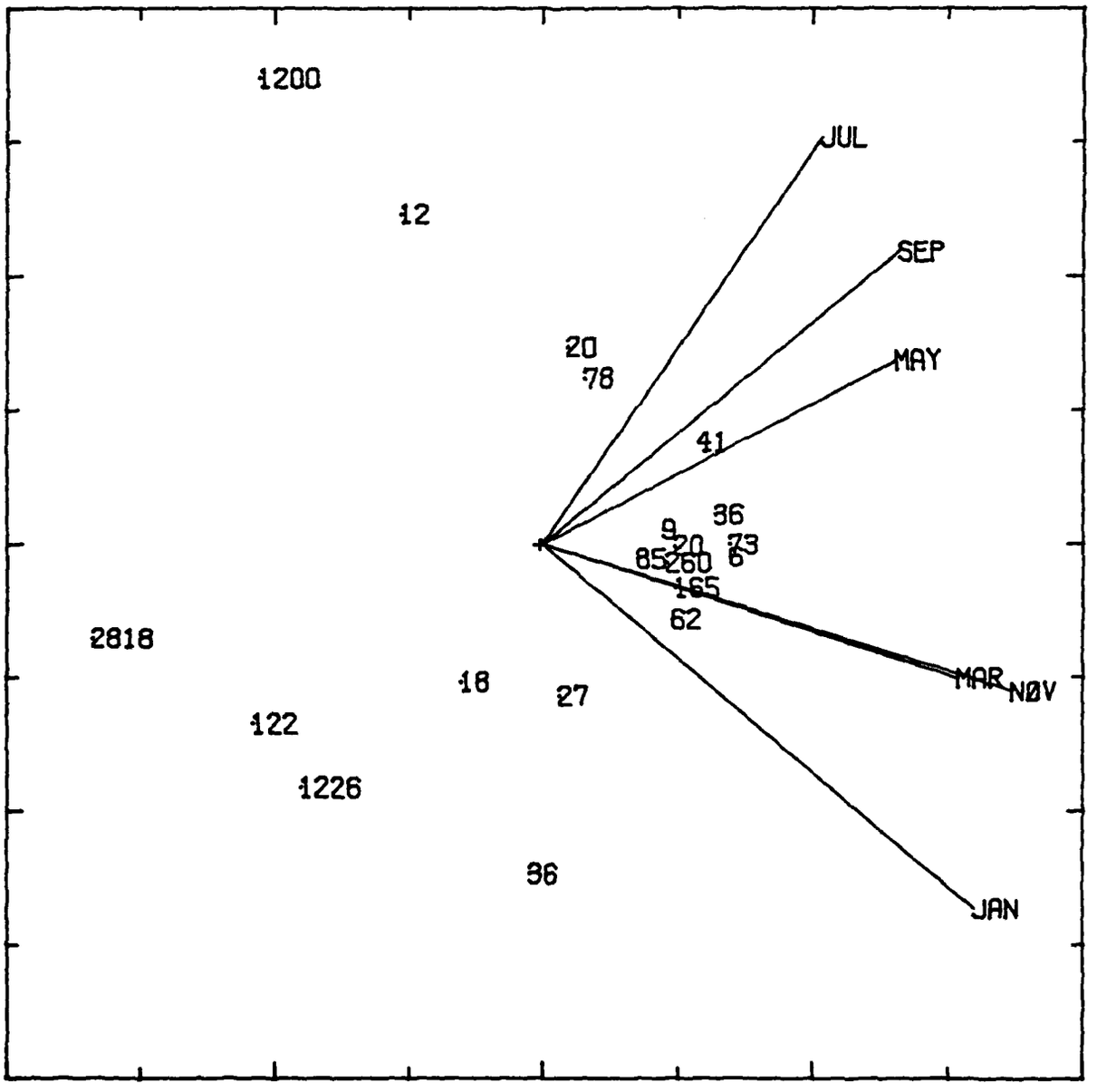
analysis are less obvious. A good idea is to record the extraneous measurements on the g points of the biplot and see if they show some regularity in the plane. Thus, in the temperature example we have marked the altitude on each g-point in Figure 7 and we see at once that the distribution on the biplot shows much regularity.

The leftmost g-points are those of stations at high elevations - evidently there is some right to left trend in altitude. As this trend is in direction opposite to the general direction of the h-arrows for months' temperatures, one would conclude (again unsurprisingly) that temperatures are rather lower at higher elevations.

Perhaps some additional comments on this example would further illustrate uses of the biplot. The North-South differences of Figure 6 and the altitude differences of Figure 7 account for a great deal of the variability of the stations. However, a number of stations do not quite fit the pattern, especially stations 2 and 10 which are much farther to the right of the biplot than one would expect from their altitudes. Checking their locations on Figure 2 one sees these stations to be far inland on the South American continent. Evidently, in addition to altitude and to Northern versus Southern latitude, distance inland also plays a role in determining temperatures.

This illustration shows how the biplot can be used to check hunches about relationships to extraneous variables and how

Figure 7: Biplot of Temperature Data With Indication of Altitude of Stations



inspection of the biplot may suggest new things to look for. Of course, these are subjective impressions and their effective use depends very much on the ideas the investigator may be able to generate. The biplot will help him; it will not provide an objective substitute for his intuition.

4.3. Clustering of units

Some groupings of units may be evident from the inter-unit distances themselves, rather than from extraneous information. Such data-dependent groups will be referred to as clusters, and methods of defining such groupings will be referred to as clustering algorithms: They differ from those used for locating sheaves of variables in that they relate to units rather than to variables and that the criterion for clustering is small inter-unit distances, whereas the criterion for forming sheaves was high inter-correlation of variables.

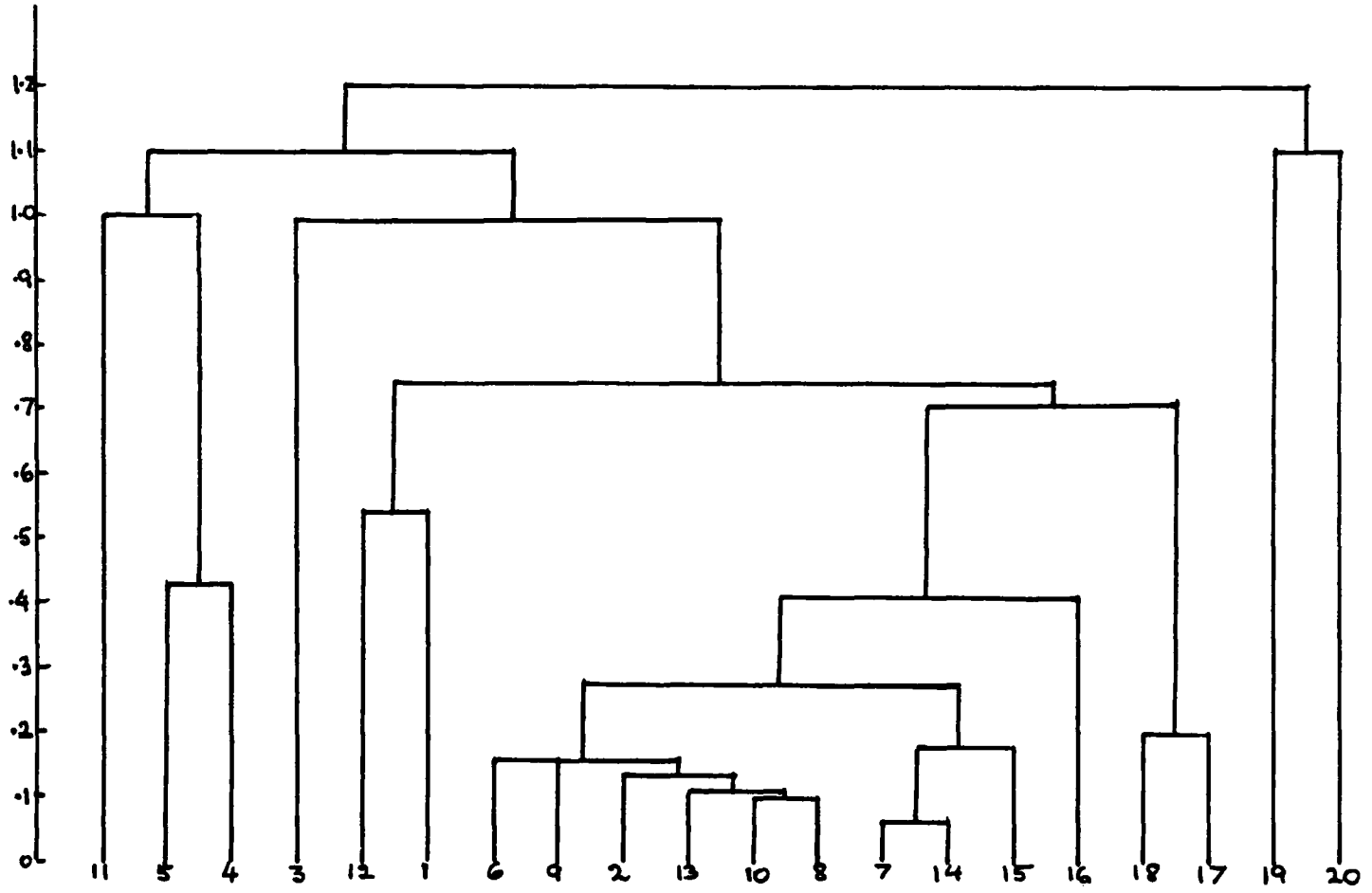
Many methods of clustering are available. A very simple one uses *single linkage*. To begin with, one clusters the nearest two points together. At the second stage one considers the next smallest distance: If it is between one of the first two points and a third point one clusters all three points together; if it is between two other points, one forms a second cluster of those two points. At each successive stage one considers the smallest of the distances betweeen points which are not already in the same cluster. The points separated by this least distance are then linked together, and with them any other points clustered previously to either of them. Thus, at a particular stage units 8 and 15 have least distance 0.28, and in previous stages unit 8 had been clustered with units 2, 6, 9, 10 and 13 whereas unit 15 had been clustered with units 7 and 14, then the new cluster consists of units 2, 6, 7, 8, 9, 10, 13, 14 and 15.

One may thus proceed step by step up to the largest distance between points, at which stage all units become a single cluster. In practice one will presumably want to stop the clustering process before that, either when the number of clusters is small enough or when the remaining distances are too large.

The entire clustering process can be displayed by a dendogram which is an inverted tree-like structure with a vertical scale corresponding to distance. This dendogram has a single stem on top at the height of the largest distance, when all units are clustered together. At the bottom of the dendogram, below the height of the least distance, it has n separate branches, one for each unit. In between, at the height of each distance, it has as many branches as there are clusters at that distance. Below that height the branch further branches and sub-branches until the individual units' branches are reached.

The nearest neighbor dendogram for the 20 stations of the temperatures example -- corresponding to the standardized biplot distances in Table 9 -- is given in Figure 8.

Figure 8: Temperature data - single linkage dendogram of biplot distances



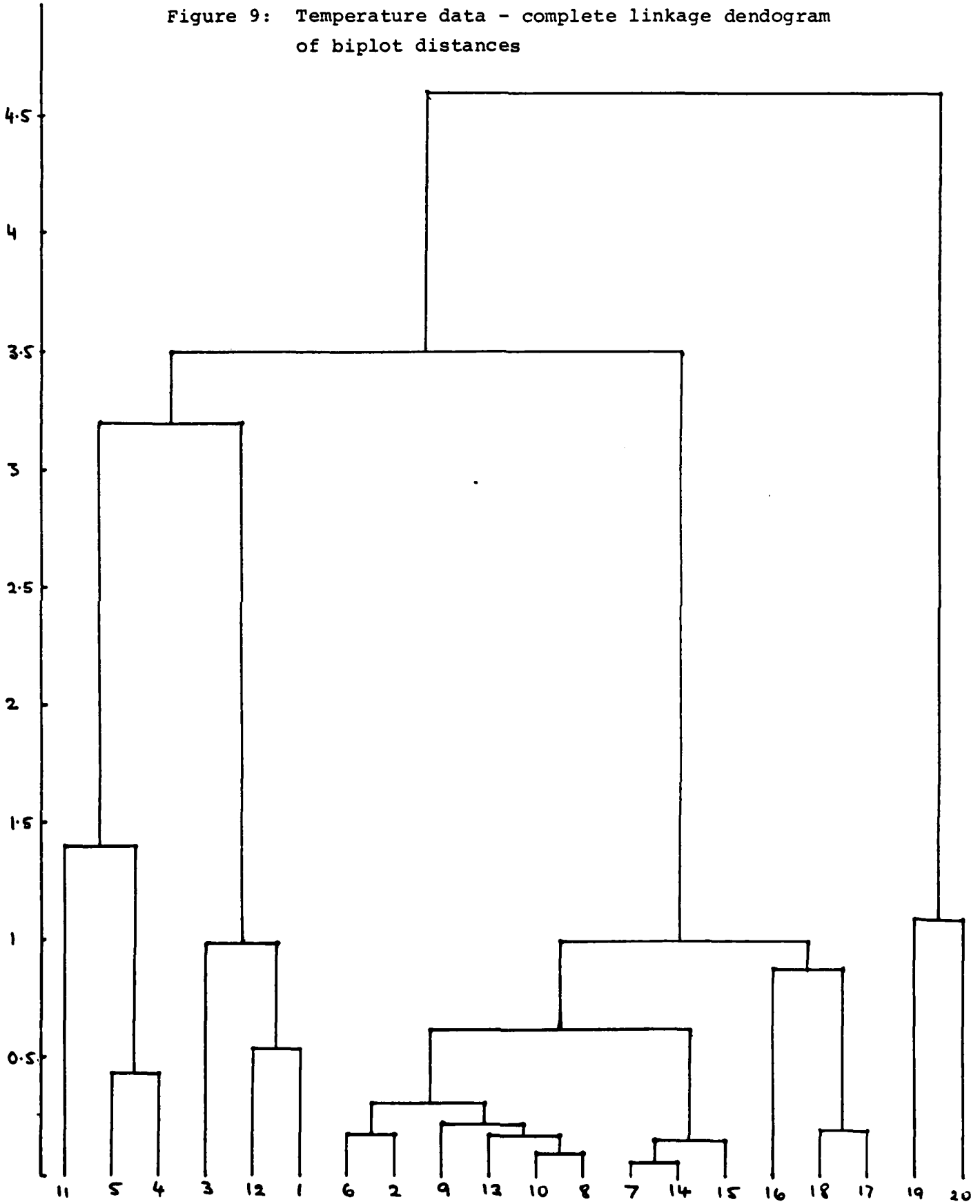
(For convenience, the order of the points has been re-arranged to correspond as closely as possible to that of the biplot -- in practical application this is of course not possible since the "true" order is not known.)

It will be seen that the dendogram of Figure 8 reproduces only some of the clusters evident from the biplot g-point scatter of Figure 2. Thus if we divide the points into four clusters by lopping off the branches between heights 1.0 and 1.1, one cluster is of stations 4, 5 and 11, two are of the single stations 19 and 20, and one of the remaining fifteen stations. This is not very satisfactory because the last cluster is too large and spread out: The largest intra-cluster distance is 2.9 between stations 8 and 17. Such an elongated "cluster" is obtained because there is a "chain" of points at relatively small (below 1.0) distances from point 3 to point 17, i.e., 3 to 1 (.99), 1 to 6 (.74), 6 to 15 (.62), 15 to 16 (.41) and 16 to 17 (.89).

An alternative clustering criterion which would avoid such "elongated" clusters uses the complete linkage and clusters a set of units together, at distance d_0 and above, only if all units within the set are within d_0 of one another. The corresponding dendogram for the temperature data is shown in Figure 9. It is seen to differ from Figure 8 not only in that it shows less clustering for each given distance, but also in that it results in somewhat different clusters.

Thus, to obtain four distinct clusters, one would lop off the branches at $d_0 = 2$ and the resulting clusters would be 5, 4, 11 (as before); 19, 20 (which had been separated before); 1, 3, 12 and the remaining twelve stations (these last two clusters formed a single cluster by the previous method). The separation of that elongated cluster into two tighter clusters seems more satisfactory.

Figure 9: Temperature data - complete linkage dendogram of biplot distances



In addition to these two clustering criteria, there are many others in the literature with algorithms and computer programs to carry them out. Each of the methods is supposedly "objective," but the choice of a method is a subjective matter, and each investigator must make sure he is using a method whose criterion is meaningful to him and appropriate to his purposes.

When the units cluster "naturally" into distinct tight groups, pretty much all clustering algorithms will reproduce that pattern. Often, however, the scatter does not reveal such obviously distinct clusters and their different algorithms will output different "clusters." In such cases one would be justified in using some "objective" method only if one were really satisfied with the relevance of its criterion. Otherwise, analysis into "clusters" becomes a game, especially if one tried out a variety of algorithms and then picked out one of them. Indeed, the multiplicity of available algorithms pretty much guarantees that any random scatter shall "cluster" nicely by some one of the many criteria. The investigator should be cautioned to inspect the clustering criterion carefully before he commits his data to an "objective" analysis into clusters.

The virtue of objectivity in data analysis is not obvious. A subjective approach which allows some capable researchers to obtain insights is certainly preferable to an objective method which usually fails to reveal anything worthwhile to any investigator. One should not carry democracy too

far. In the analysis of scatters of units (as well as that of configurations of variables), the capable investigator will usually approach his data with a great deal of prior knowledge, hunches and hypotheses about patterns and relationships. He will do well to be guided by them and direct his analysis accordingly. If he wishes to cluster his data, he should not do so "objectively," merely on the basis of distance (or correlations) but should allow the interplay of observation with prior hypothesis. Specifically, if unit U_1 is about as distant from unit U_2 as from unit U_3, the investigator would do well to group it with the unit with whom he has a priori reason to expect it to be more closely related.

Clustering algorithms are popular not only because they are "objective" (after subjective choice of the algorithm) but because they can deal with large scatters (or configurations) and have been programmed. It is very difficult to inspect large data matrices by eye, though use of prior ideas about possible patterns may be of great help. (See for example Guttman's use of linear and circular dependence patterns -- the simplex and radex (Guttman, 1954) -- for meaningful inspection of correlation matrices.)

4.4. Outliers

Distances $\sqrt{d_{i,i'}}$, are standardized by definition, (Section 1.3, above). As a result, the scatter of points in m-space is spherically symmetric, and its approximation on the biplot is essentially circular. Unlike the well-known elliptic forms of variability of row variables, their standardized representation is circular.

In studying the form of the distribution, therefore, we should not look for asymmetry -- which has been eliminated by standardization -- but rather for other features such as clumping or clustering of points, special patterns, associations with external variables, outliers, etc.

To begin with, note that the sum of squares of standardized distances is fixed by standardization. Thus,

$$\sum_{i=1}^{n} u_{i,i} = m \qquad (4.1)$$

so that the average of the squared distances from the centroid is

$$\bar{u} = m/n. \qquad (4.2)$$

Also, by the triangular inequality for distances one obtains

$$\sqrt{d_{i,i'}} \leq \sqrt{u_{i,i}} + \sqrt{u_{i',i'}} \quad . \qquad (4.3)$$

For the biplot approximations these correpsond to

$$\Sigma_i ||\underline{g}_i||^2 = m, \qquad (4.4)$$

and

$$||\underline{g}_i - \underline{g}_e|| \leq ||\underline{g}_i|| + ||\underline{g}_e|| \ . \qquad (4.5)$$

If the scatter is roughly evenly distributed within radius unity, there is little to be said. If, however, one notes isolated points in one direction, with the remaining $\underline{g}$'s tightly bunched in the opposite direction, one should inspect the outlying points carefully for measurement or recording errors or perhaps for not belonging to the population under study. If so, one might do well to omit such units from analysis and concentrate on the units that have a reasonable statistical scatter. This would, of course, mean recalculating the principal axes and $\underline{g}$ and $\underline{h}$ vectors after omission of the outliers.

A reasonable criterion for multivariate outliers is the distance $\sqrt{u_{i,i}}$ from the centroid. One does well to look at the distribution of these n distances and see if it indicates some clearly outlying units. Tests of significance are available for the multi-normal case (Gnanadesikan, 1977) but we feel that these should be used with great caution unless one really has good reason to believe that the data came from such a distribution.

It often happens that one finds one or more "outliers" but checks do not reveal any reason why those observations should be unusual. So one does not know whether these are extreme values which do occur sometimes, though rarely, in the given observational situation, or whether these are erroneous records which do not belong with the batch under study. One is in a quandary as to whether to "reject" such outliers or not. Not to reject means including observations that manifestly do not fit the statistical distribution of the majority and vitiates the assumptions underlying most statistical procedures. To reject exposes one to risks of biasing the statistical analysis if the outliers were extremes from the same distribution as the rest of the observations. An honest rule would be always to report at least the number of rejected outliers and preferably their entire observations, but not to include them in the main statistical analysis. Such rules and considerations apply as much to multivariate as to univariate data.

4.5. The distribution of points in the g scatter

Some idea of the distribution of the batch over the variables may be obtained from considering the scatter of $\underline{g}$-points on the biplot. A more or less regular unimodal distribution should result in a reasonably symmetrical biplot scatter with a concentration of points about the centroid and gradual tapering off density towards the edges.

Some other distributions obviously have different biplot scatters. A common case is that of multivariate J-shaped distributions which have a mode near zero for all variables and a density which decreases for higher values of each variable. Such distributions will produce biplots of the type illustrated in Figure 10 -- essentially a quadrant of points with a high concentration at the vertex and along the edges and with $\underline{h}$-arrows pointing in the direction opposite to the vertex. The vertex represents the zero point of all variables, the edges the zeroes of particular variables.

To check such distributions it is useful to project individual numerical vectors, as in particular the zero vector, onto the biplot scatter. For the zero $\underline{z}_0 = \underline{0}$, the projection is

$$\underline{g}_0' = -\underline{\bar{z}}'F' \qquad (4.6)$$

as in (2.12). The rough location of the zero vector is indicated on Figure 10 and confirms the supposition that these data are of a multivariate J-shaped distribution.

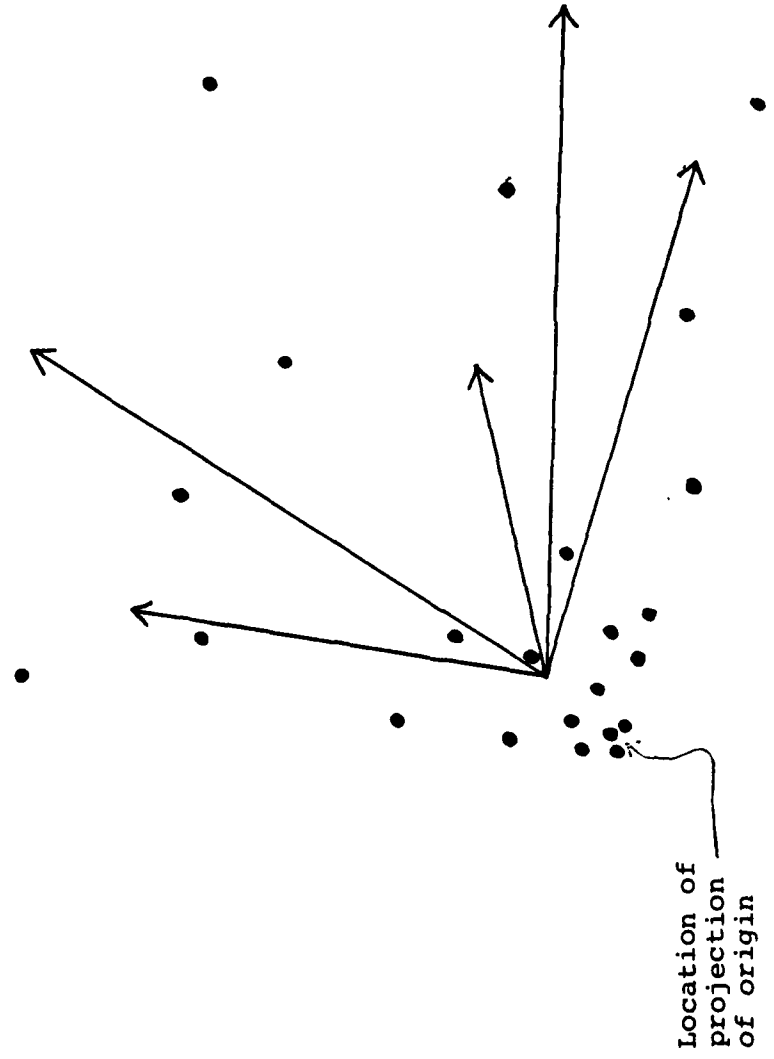


Figure 10: Biplot of hypothetical data with 5 J-distributed variables

When the g-scatter does not show any special pattern one must consider the distribution to be essentially random. It is not that regularities may not exist, but that they are not evident from the scatter. We know of no way of "testing" for normality, and would rather tend to use the normal model by default, as a viable model in case nothing contrary emerges from biplot inspection.

In effect, the biplot may provide a more sensitive check of multivariate normality than the commonly available tests of significance which concentrate on each individual variable rather than on a plane as the biplot does. But, as of now we have no way of using the biplot plane for a test of significance on the shape of distribution.

Another intriguing issue is whether the biplot might be suggestive of a transformation to normality. All we can say at this time is that strongly skewed distributions should show up in a biplot looking like that of Figure 10. Hence, the appearance of such a scatter might be suggestive of a transformation by square roots, logarithms or similar functions. This subject needs further examination.

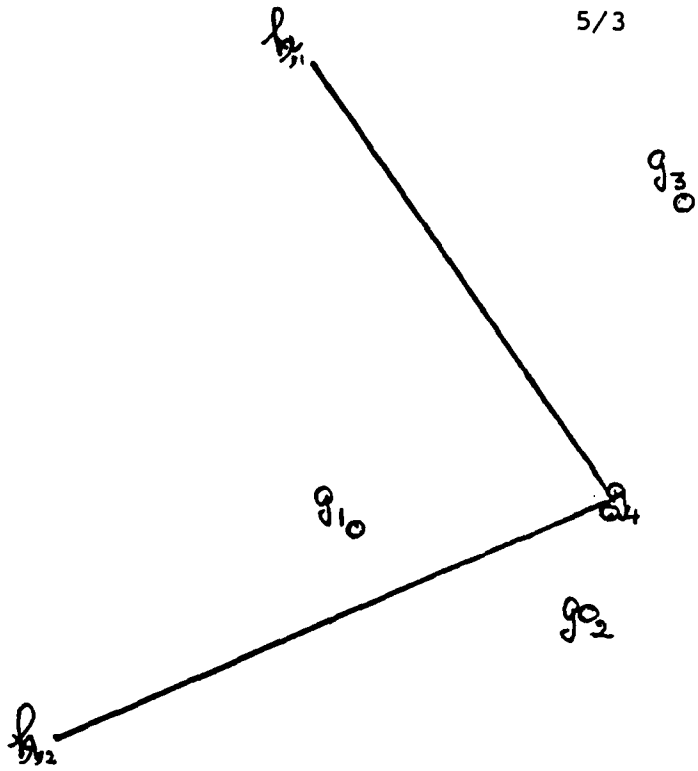
5. JOINT ANALYSIS OF VARIABLES AND UNITS - MODELLING

5.1. Importance of joint display in the biplot

The biplot jointly displays the configuration of the variables (columns of the data matrix) and the standardized scatter of the units (rows of the data matrix). In doing both these things simultaneously, it differs from many other displays, which concentrate on one feature to the exclusion of the other. Multidimensional scaling models either the correlation matrix of the variables or a distance matrix for the units, but not both. It is not usually feasible to bring in the variables into the multidimensional scaling of units, or vice versa. (See, however, Gabriel, 1978). As a result, the analysis and interpretation provided by such scaling is more limited than that provided by biplot representation. Multidimensional scales may have more flexible fitting algorithms and are not restricted by the geometry of least squares, but they are more limited in what they display.

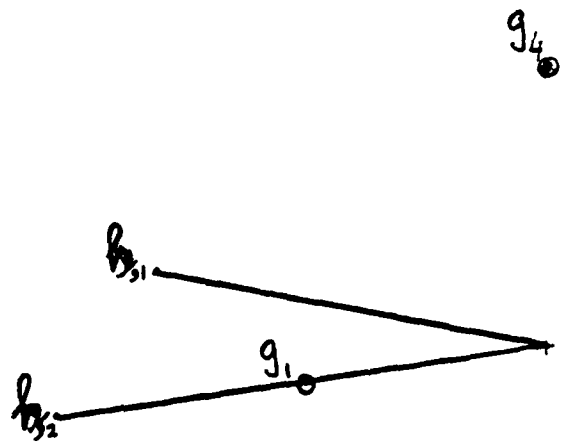
Some of the uses of the biplot in interpreting units' clusters in terms of variables have been discussed above. Analogously, correlations can sometimes be explained in terms of the scatter of units -- in particular, sometimes a single outlier in a particular direction can account for an increased correlation of the variables displayed in its direction. This is illustrated by the two parts of Figure 11.

Figure 11a shows the biplot of a 5-observation 2-variate matrix: the low correlation (r=0.20) is evident in the close to right angle (78°) between the two $\underline{h}$ vectors. The scatter of points, however, shows $\underline{g}_3$ to be fairly separate from $\underline{g}_1$, $\underline{g}_2$, $\underline{g}_4$ and $\underline{g}_5$ -- evidently an outlier. Removal of the third point leads to a new configuration, biplotted in Figure 11b, and demonstrating a much higher correlation (r=.94) as evident from the 20° angle between $\underline{h}_1$ and $\underline{h}_2$.



$$Y = Z = \begin{pmatrix} 11 & 25 \\ -7 & 7 \\ 18 & -18 \\ -1 & 1 \\ -21 & -15 \end{pmatrix}$$

Fig. 11a A biplot with an outlier



$$Z = \begin{pmatrix} 11 & 25 \\ -7 & 7 \\ -1 & 1 \\ -21 & -15 \end{pmatrix}$$

$$Y = \begin{pmatrix} 15.5 & 20.5 \\ -2.5 & 2.5 \\ 3.5 & -3.5 \\ -16.5 & -19.5 \end{pmatrix}$$

Fig. 11b The biplot after removal of the outlier

5.2. Diagnosing models by means of the biplot

Approximate functional fits of the data matrix may at times be identified by inspection of the biplot. Thus, if Z is approximately additive, i.e., if

$$z_{i,v} = \bar{\bar{z}} + a_i + b_v + e_{i,v} \qquad (5.1)$$

for

$$\bar{\bar{z}} = \Sigma_i \Sigma_v \, z_{i,v}/nm \qquad (5.2)$$

and some $a_1,\ldots,a_n$, $b_1,\ldots,b_n$ and small e's, then the biplot of Z -- or of $((z_{i,v} - \bar{\bar{z}}))$ -- will have the following simple form: The g-markers will be close to one straight line, the h-markers close to another such line and these two lines will be at 90° to each other. Conversely, when the biplot markers display such a pattern, additivity can be inferred.

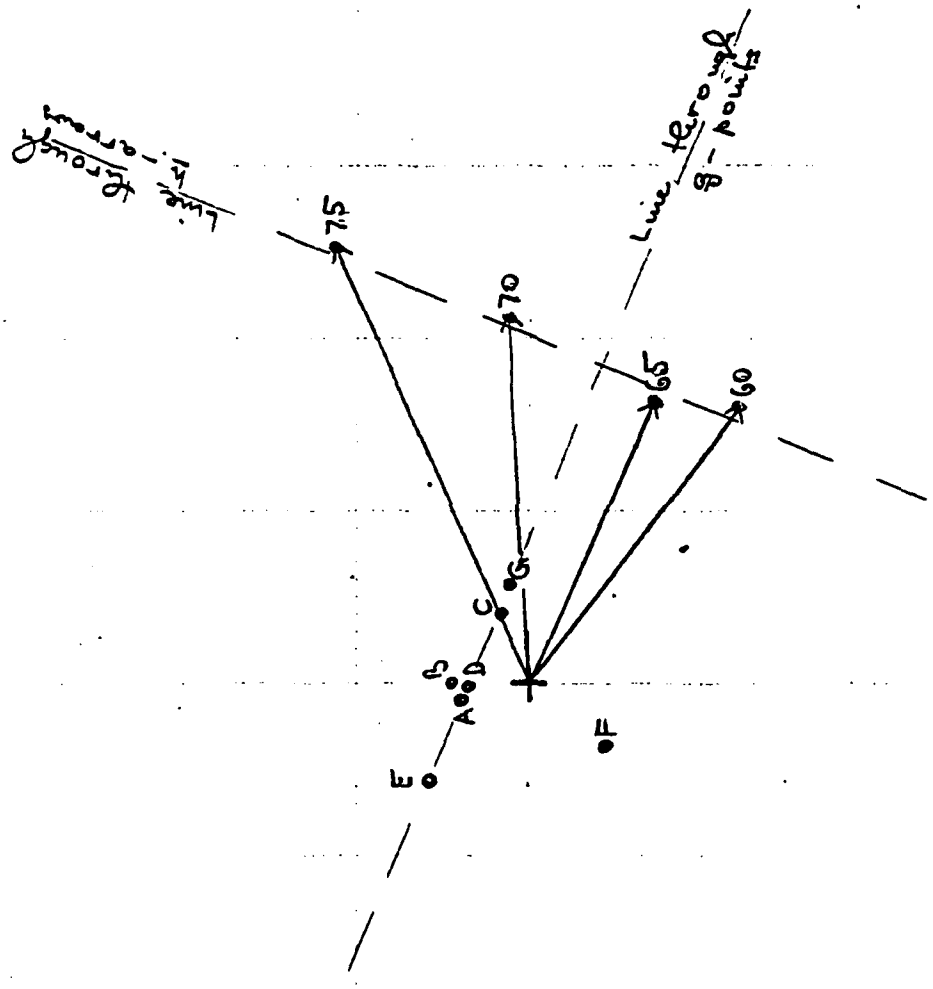
What is more, if some row markers are on one line and some column markers are on another line which is at 90° to the first, then one can infer that additivity holds for the sub-matrix of the corresponding rows and columns. For an illustration, consider the artificial air pollution data of Table 13 which is biplotted in Figure 12. It is immediately evident that the heads of the h arrows for the four years are very close to collinear and that the g-point for six of the stations are close to another line, pretty much at 90° to the h-arrowhead line -- only the g-point for station F is

TABLE 13

An Air Pollution Index at Seven Localities 1960-75 (Artificial Data)

Station	1960	1965	1970	1975
A	100	102	105	110
B	98	99	104	108
C	107	110	112	116
D	98	100	103	106
E	86	90	91	95
F	103	100	94	89
G	111	111	115	119

Figure 12 Biplot of Ficticious Air Pollution Data



far away from this line. One may therefore safely diagnose an additive model for the 6 x 4 table obtained by omitting station F. Inspection of the Table will show that this is indeed appropriate.

This diagnostic method extends to some other models as well (Bradu and Gabriel, 1977). In particular, if the two above lines intersect at an angle other than 90°, then a Tukey degree-of-freedom-for-non-additivity model holds, i.e.,

$$z_{i,v} = \bar{\bar{z}} + a_i + b_v + \lambda a_i b_v + e_{i,v} \qquad (5.3)$$

for some λ.

This diagnostic use of the biplot may be quite important since statisticians do not in general have adequate tools for such diagnosis. Statistics textbooks generally give methods of estimating parameters and testing fit of given models, but do not usually provide techniques of choosing a model.

Biplot diagnosis of models rests on the matrix decomposition

$$Y \underline{\underline{apx}} GH' \qquad (5.4)$$

The rows of the latter two matrices are displayed in the biplot where visual inspection may lead to diagnoses of simple geometric descriptions. When these descriptions are formulated algebraically they can be entered into (5.4) and may be translated into a model for the data matrix itself.

5.3. An example of modelling by means of the biplot

As an example of the biplot's usefulness in modelling, consider the case where the vertices of the $\underline{h}$-arrows are close to an ellipse. Writing $\underline{\mu}$ for the center of the ellipse, $\underline{\alpha}$ and $\underline{\beta}$ for unit vectors along its principal axes, this means that there exists θ_v for each v, such that

$$\underline{h}_v = \underline{\mu} + \underline{\alpha} \cos \theta_v + \underline{\beta} \sin \theta_v \quad . \tag{5.4}$$

Matrix H therefore becomes

$$H' = (\underline{\mu},\underline{\alpha},\underline{\beta}) \begin{pmatrix} 1 & \ldots & 1 & \ldots & 1 \\ \cos\theta_1 & \ldots & \cos\theta_v & \ldots & \cos\theta_m \\ \sin\theta_1 & \ldots & \sin\theta_v & \ldots & \sin\theta_m \end{pmatrix} \tag{5.5}$$

and the data matrix is approximated by

$$Y \underline{\underline{\text{apx}}}\ G(\underline{\mu},\underline{\alpha},\underline{\beta}) \begin{pmatrix} \ldots & 1 & \ldots \\ \ldots & \cos\theta_v & \ldots \\ \ldots & \sin\theta_v & \ldots \end{pmatrix} \quad . \tag{5.6}$$

Thus, the i-th row is approximated by

$$\underline{y}_i' \underline{\underline{\text{apx}}}\ (g_i'\underline{\mu},\ g_i'\underline{\alpha},\ g_i'\underline{\mu}) \begin{pmatrix} \ldots & 1 & \ldots \\ \ldots & \cos\theta_v & \ldots \\ \ldots & \sin\theta_v & \ldots \end{pmatrix} \quad . \tag{5.7}$$

Writing

$$\sin \phi_i = \frac{\gamma_i}{\sqrt{\gamma_i^2 + \delta_i^2}} \quad ,$$

$$\cos \phi_i = \frac{\delta_i}{\sqrt{\gamma^2 + \delta^2}} \quad ,$$

and

$$\mu_i = \underline{g}_i' \underline{\mu} \quad , \tag{5.8a}$$

$$\gamma_i = \underline{g}_i' \underline{\alpha} \quad , \tag{5.8b}$$

$$\delta_i = \underline{g}_i' \underline{\beta} \quad . \tag{5.8c}$$

These approximations become

$$y_{i,v} \underline{\underline{\text{apx}}} \; \eta_i + \gamma_i \sin\theta_v + \delta_i \cos\theta_v \quad , \tag{5.9}$$

or, defining

$$\Psi_i = \sqrt{\gamma_i^2 + \delta_i^2} \tag{5.10}$$

and $$\Phi_i = \arctan (\gamma_i/\delta_i) \tag{5.11}$$

they become $$y_{i,v} \underline{\underline{\text{apx}}} \; \eta_i + \Psi_i \cos(\theta_v - \Phi_i) . \tag{5.12}$$

Thus, observation of the elliptical form of the $\underline{h}$-configuration has led to diagnosing a harmonic model for the data matrix with constant and amplitude depending on the rows and phase on the columns. An example where

such a model was appropriate was given in Section 2.6 where the elliptical configuration of the months' arrows was found. Those annual temperature data could therefore be fitted by a harmonic model with constants and amplitudes depending on the station and the phase depending on the month (Gabriel and Tsianco, 1980).

Let it be stressed that the function of the biplot, or bimodel, is merely to suggest a suitable model, not to provide estimates of its parameters. Once a model such as (5.1), (5.3) or (5.11) is suggested, standard estimation techniques should be reverted to, such as least squares or its robust counterparts. (We will not discuss the fitting of the harmonic model here -- see, however, Gabriel and Tsianco, 1980).

6. COMPARING SEVERAL BATCHES OF OBSERVATIONS

6.1. Joint inspection of the two batches' scatters

Observations coming from several different sources, or populations, need different methodology and analysis than single batches of multivariate data. For each single batch of data, one may be concerned with description and analysis of the configuration of variables and of the scatter of units and with consideration of distributions, outliers, models, and other summarizations. This, of course, may also be of interest when several batches of data are available, but the new aspect that appears at this stage is that of comparing batches. Problems now arise with the search for, and identification of, characteristics on which the batches differ, with the measurement of "distances" between batches, with the appraisal of the significance or possible randomness of observed differences and with the classification of additional units as being similar to one or another of the batches according to these units' multivariate observations.

One may begin with the most straightforward case of two batches. In comparing two batches one generally ignores the individuality of the units and regards them as mere members of one batch or the other. Since each batch is regarded as a sample from some population or distribution the units lose their identities and become mere replicate observations.

In comparing two samples, and in using them for testing population differences, one considers the within sample, inter-unit differences, mainly as providing estimates of random variation, or "noise," against which to judge inter-sample differences (averaged over units). Thus, in a comparison of 1977 winter storms with 1978 winter storms, the individual storms of each year are averaged for the main comparison, and the variability from storm to storm within each year serves as a yardstick against which one may measure the averages' comparison. A study of the special features of individual storms of either season would be part of each batch's analysis, not part of the batch-to-batch comparison.

As an example, consider the data in Table 14 relating to 26 storms occurring in the summer of 1973. Pielke and Biondini (1977) treated these as two batches of storms, 13 with geostrophic wind speed above 3m/sec and 13 with slower geostrophic wind speed. In comparing these two batches, the individual storms are averaged for comparison, and the storm to storm variation within each batch provides estimates of random variability. A study of the special features of each batch's individual storms is not a main part of the batch-to-batch comparison. Table 15 gives the five-variate means of each batch and the variance-covariance estimates from each batch.

TABLE 14

Summer 1973 Storms

Date	R	Wind Speed	D Direction	T	S	P
July 3	49.61	4m sec^{-1}	90°	.019	5.99	432
July 4	172.83	2m sec^{-1}	70°	.032	5.98	453
July 5	20.72	2m sec^{-1}	225°	.054	5.77	371
July 6	59.93	4m sec^{-1}	270°	.032	10.77	515
July 7	26.12	3m sec^{-1}	135°	.048	11.58	494
July 15	75.48	3m sec^{-1}	100°	.082	8.41	336
July 16	51.71	5m sec^{-1}	170°	.068	14.14	267
July 17	56.33	6m sec^{-1}	100°	.042	10.57	477
July 18	23.66	5m sec^{-1}	100°	.081	13.36	326
July 20	62.95	3m sec^{-1}	135°	.076	12.70	357
July 23	31.13	6m sec^{-1}	120°	.048	9.35	539
July 24	17.09	10m sec^{-1}	90°	.096	10.44	404
July 25	14.61	6m sec^{-1}	85°	.081	3.64	356
July 26	37.29	2.5m sec^{-1}	110°	.074	11.57	304
July 27	84.05	1m sec^{-1}	180°	.047	11.04	316
July 28	77.22	1m sec^{-1}	170°	.039	9.22	329
July 29	108.71	0	180°	.053	12.97	295
July 31	93.88	6m sec^{-1}	180°	.120	16.59	280
Aug 1	38.66	1m sec^{-1}	180°	.116	15.53	252
Aug 2	75.61	3m sec^{-1}	190°	.113	12.54	242
Aug 6	79.98	4m sec^{-1}	100°	.070	15.53	317
Aug 10	127.04	3m sec^{-1}	140°	.028	4.46	564
Aug 11	24.85	4m sec^{-1}	135°	.058	9.59	401
Aug 12	17.66	8m sec^{-1}	90°	.081	15.85	427
Aug 13	33.15	7m sec^{-1}	100°	.043	9.19	338
Aug 14	97.53	3m sec^{-1}	135°	.050	7.80	532

Fast: surface geostrophic wind speed $\geq$ 3m/sec; slow: other

R: rainfall 20 log

D: surface level geostrophic wind direction

T: gradient of equivalent potential temperature

S: difference between saturation equivalent potential temperature and equivalent potential temperature

P: depth of convective instability

TABLE 15

Means and Variances-Covariances of Storm Data (Transformed)*

Fast Geostrophic Wind (Speed > 3m/sec) - 13 storms

		R	D	T	S	P
Means		215.10	125.38	248.03	111.55	294.82
St. Devs.		35.17	51.12	54.96	37.03	31.43
Variances - Covariances	R	1237.06	910.08	-405.95	540.62	-194.38
(Correlations below	D	.506	2613.31	-327.15	470.87	139.69
diagonal)	T	-.210	-.116	3020.83	1032.91	-999.24
	S	.415	.249	.408	1370.91	-459.90
	P	-.176	.087	-.579	-.395	987.61

Slow Geostrophic Wind (Speed ≤ 3m/sec) - 13 storms

		R	D	T	S	P
Means		251.09	150.00	244.35	99.67	287.04
St. Devs.		36.05	40.76	52.47	31.85	38.27
Variances - Covariances	R	1299.37	-603.74	-719.94	-330.53	296.38
(Correlations below	D	-.411	1661.54	499.50	321.12	-649.04
diagonal)	T	-.381	.234	2752.82	1183.67	-1472.29
	S	-.288	.247	.708	1014.57	-855.09
	P	.215	-.416	-.733	-.701	1646.74

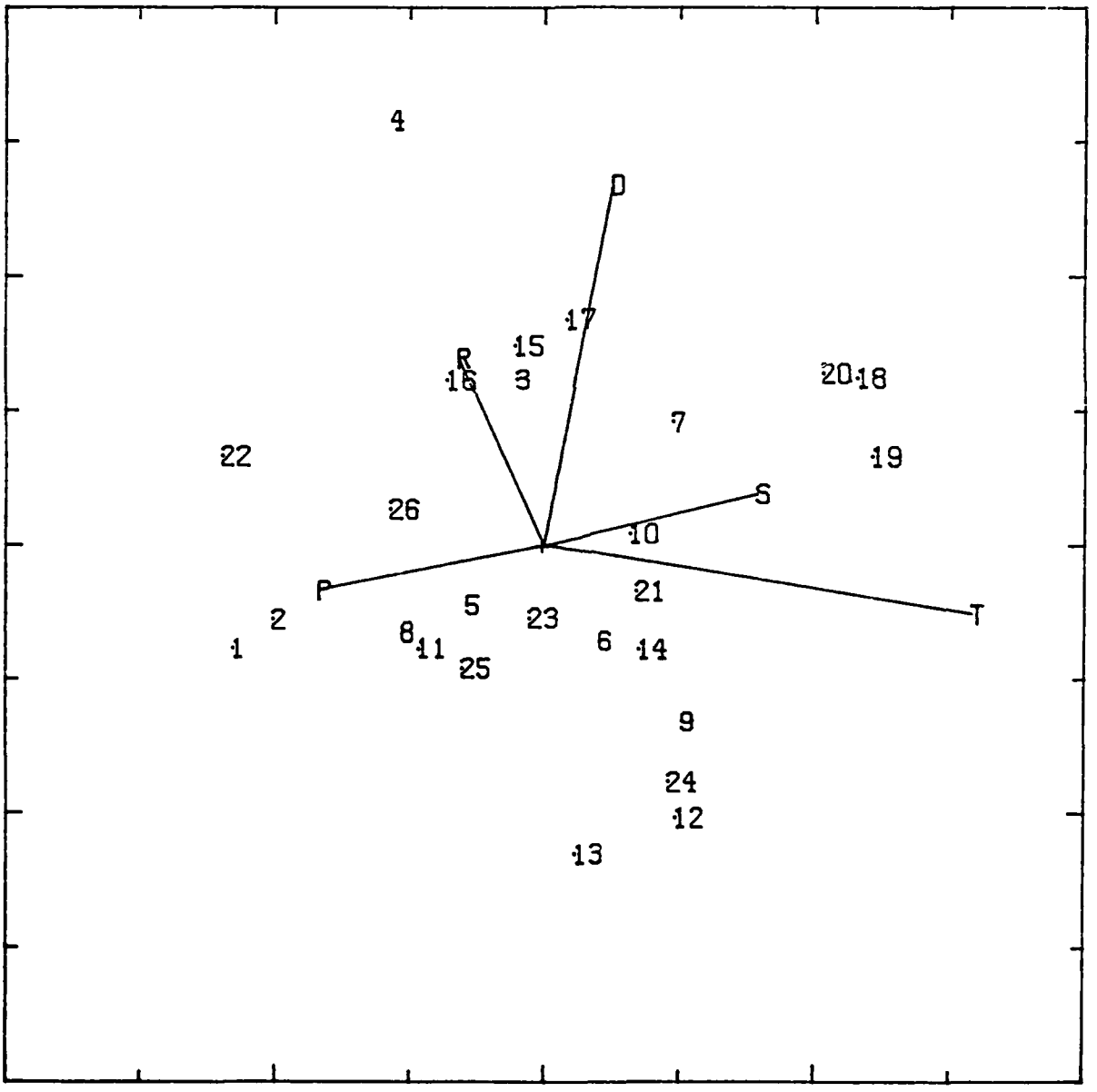
*Transformations: $R \leftarrow 60 \ln R$; $D \leftarrow D$; $T \leftarrow 1000\sqrt{T}$; $S \leftarrow 10S$; $P \leftarrow 15\sqrt{P}$.

Note that the means and variances in Table 15 do not relate directly to the variables in the form computed by Pielke and Biondini -- Table 14 -- but to various transforms of these. A preliminary check by means of probability plots showed three of the variables to have very skew distributions, especially the first one. Transformation by a fractional power was therefore indicated. After some trial and error, transformations were chosen which produced reasonably symmetric distributions. (The constants by which the variables are multiplied were chosen so as to approximately equalize the variances -- this is important for biplotting: if the biplot were fitted to non-standardized variables, the method of least squares would produce a good fit for the variables of large magnitude and all but ignore the variables of smaller magnitude.)

Such preliminary inspection and transformation of variables is quite important. Without it one might apply least squares methods to variables which are highly skewed and for which these methods would be quite unsuitable.

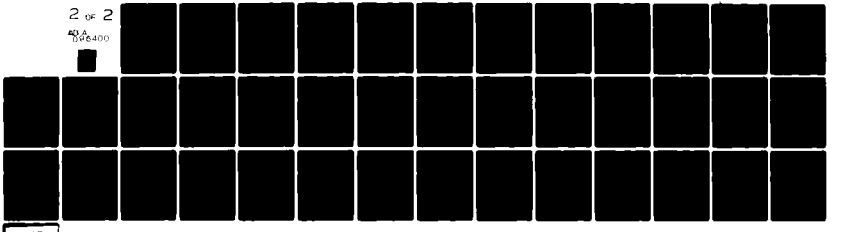
One way of representing two batches of multivariate observations is by regarding them as distinct scatters of units in the same space of variables. An approximating display -- GH' biplot -- may be constructed for the matrix of both batches' multivariate observation and the g-points of the two batches may be distinguished on the biplot by some special marks or colors. The summer 1973 storms are biplotted accordingly in Figure 13 -- again using the data transformed as noted in Table 15.

Figure 13: Biplot of Storm Data (Table 14)



AD-A096 400 ROCHESTER UNIV N Y F/6 12/1
EXPLORATORY MULTIVARIATE ANALYSIS. A GRAPHICAL APPROACH.(U)
JAN 81 K R GABRIEL N00014-80-C-0387
UNCLASSIFIED TR-81/1 NL

2 OF 2
AD A 096400



END
DATE FILMED
4-81
DTIC

It is immediately evident from this biplot that the scatters of the two types of storm differ. The most obvious difference is that the g-points for "fast" storms are mostly higher up on the biplot than the g-points for the "slow" storms. The vertical direction is that of the two variables: R-rainfall and D-wind direction. Evidently slow storms have more rainfall and higher angle of wind direction than fast storms. The two superimposed batch scatters are examined for differences in distributions. If the two scatters are completely disjoint, one may be sure of clear between-sample difference. If there is some overlap, the distinction is less obvious and may need testing for significance -- more about that later. At this stage, one does well to inspect the shape of the two scatters as well as their approximate location. If the centroids differ, this indicates a difference in mean level of some variables; the particular variables can be identified by considering the vector from one of the centroids to the other and projecting it onto the h-arrows to look for long intercepts. If the extent or shape of the two scatters differs, this indicates different variability; the particular variables on which the variability differs are indicated by identifying the h-arrows in the direction of differing scatter.

An aid in inspecting and comparing the scatter of samples of points is the construction of concentration ellipses. For a batch of m units whose biplot g-points coordinates are $(g_{i,1}, g_{i,2})$, i=1,...m, the concentration ellipse is defined as the locus of

$$\bar{g} + \beta_1 \cos\theta + \beta_2 \sin\theta \qquad 0 \leq \theta \leq 2\pi \qquad (6.1)$$

with center at point

$$\bar{\underline{g}}' = (1/m) \sum_{i=1}^{m} (\underline{g}_{i,1}, g_{i,2}) \qquad (6.2)$$

and $\underline{\beta}_1$ and $\underline{\beta}_2$ being obtained as follows: Calculate matrix

$$V = \frac{1}{m} \begin{pmatrix} \sum_i g_{i,1}^2 & \sum_i g_{i1}g_{i2} \\ \sum_i g_{i1}g_{i2} & \sum_i g_{i2}^2 \end{pmatrix} - \begin{pmatrix} \bar{g}_1^2 & \bar{g}_1\bar{g}_2 \\ \bar{g}_1\bar{g}_2 & \bar{g}_2^2 \end{pmatrix} , \qquad (6.3)$$

solve

$$V\underline{q}_\nu = \lambda_\nu^2 \underline{q}_\nu \qquad (\nu=1,2) \qquad (6.4)$$

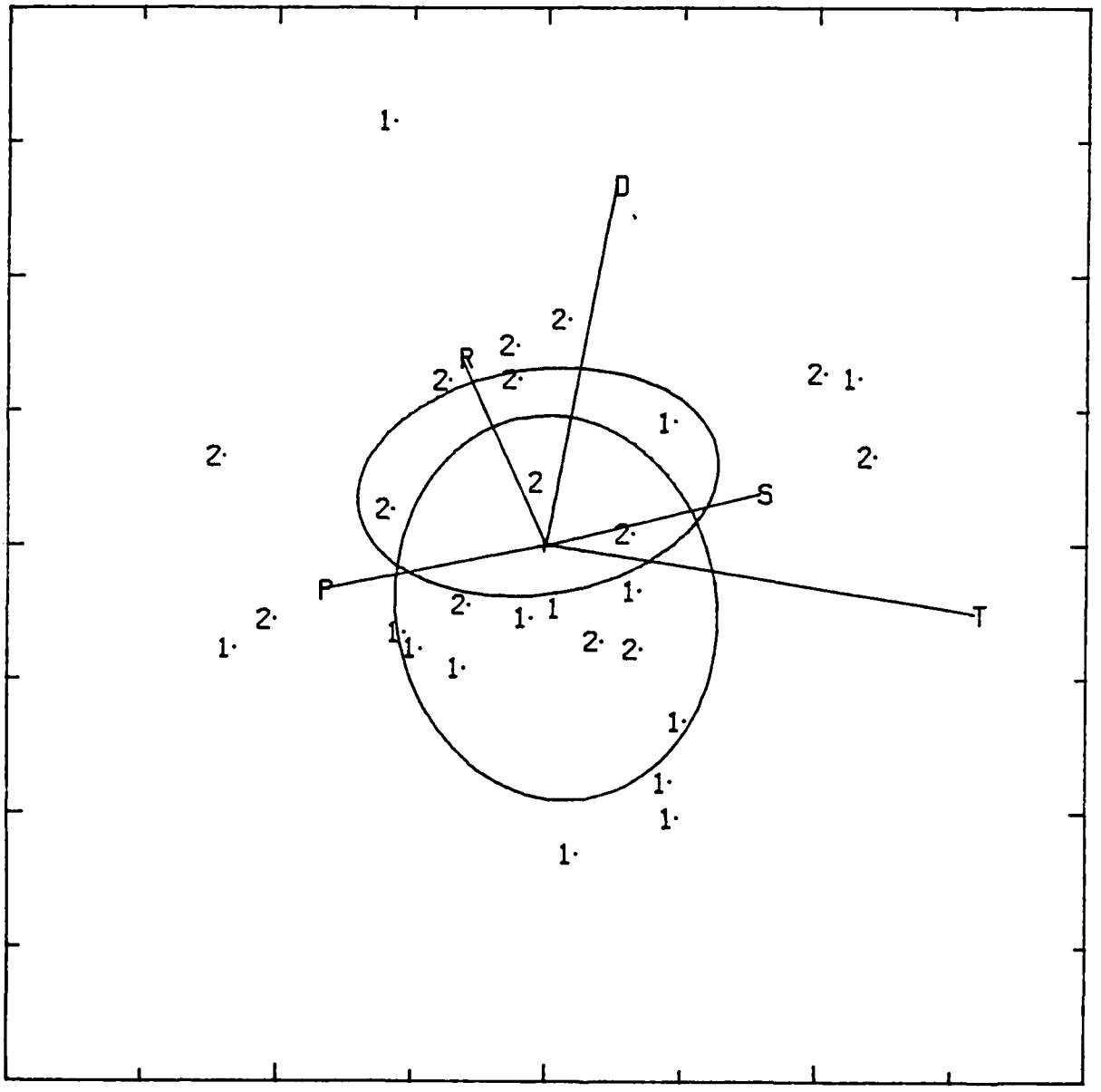
for the maximum and minimum eigen-values λ_1^2 and λ_2^2, respectively, and set

$$\underline{\beta}_\nu = \lambda_\nu \underline{q}_\nu \qquad (\nu=1,2) \quad . \qquad (6.5)$$

The center of the ellipse -- and the centroid of the batch of n $\underline{g}_i$ points -- is at $\bar{\underline{g}}'$ and the maximum and minimum diameters are, respectively, of lengths $2\lambda_1$ and $2\lambda_2$ and in directions $\underline{q}_1$ and $\underline{q}_2$ from the centroid.

The concentration ellipses for storms of each type are drawn onto the biplot in Figure 14. They clearly show the vertical displacement of the two samples, confirming the impression gained from inspection of the $\underline{g}$-points themselves. They also indicate no horizontal displacement, confirming that the two types of storms do not differ appreciably on variables T, S, and P. (The correctness of these graphical impressions can be verified from the means in Table 15.)

Figure 14: Biplot of Storms With Identification of Fast (1) and Slow (2) Storms and Concentration Ellipses

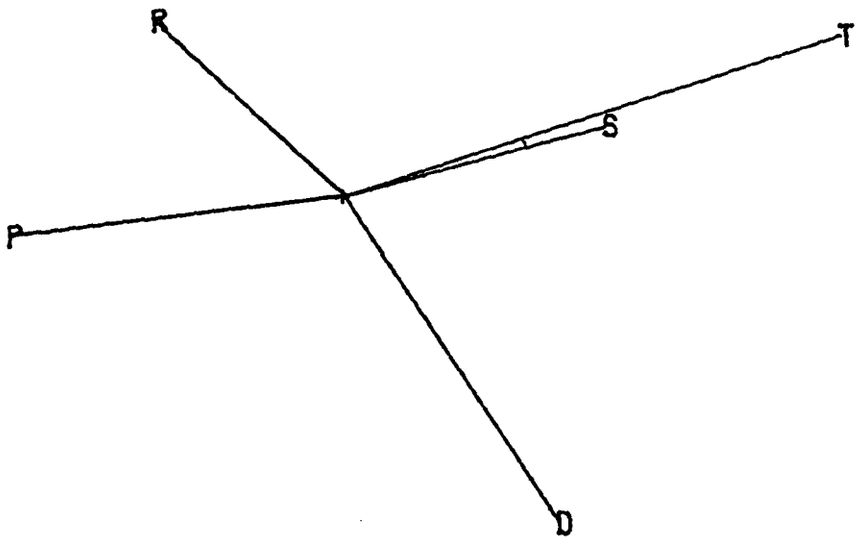


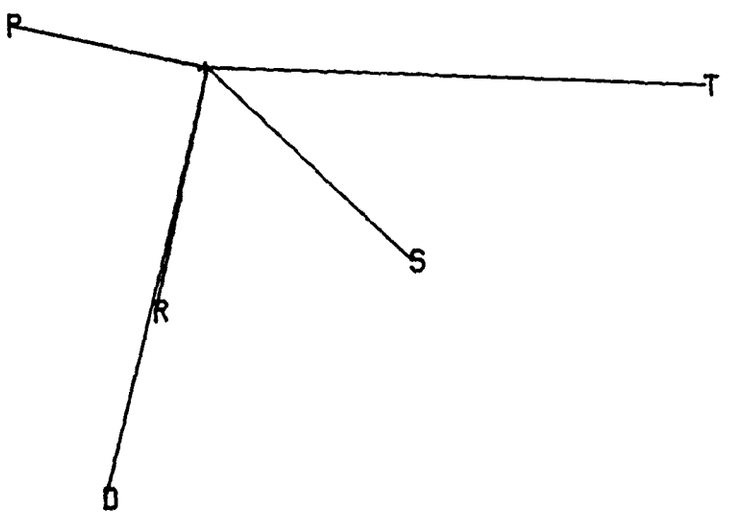
6.2. Comparison of two batches' configurations

In addition to this comparison of centroids, one may use the shape of the concentration ellipses to compare the variance and correlation configurations of the two batches. In Figure 14 the ellipse for the slow storms is considerably squatter and slightly wider than that of the fast storms. Recalling the relation of scatters to correlations, one may infer that the T, S and P correlations would not differ much between the batches, but that the correlations of R and D with each other and with T, S and P might differ. The biplot suggests that the R, D correlation is higher for fast storms than for slow ones -- and that indeed is the most striking difference between the two correlation matrices in Table 15. It also suggests that both R and D's correlations with T, P and δ are smaller in magnitude for fast storms than for slow storms, though the signs remain the same. This does not clearly reflect the actual correlations in Table 14. Evidently, comparison of ellipses can be used to indicate the existence of differences in variances and correlations, but it is difficult to use it to infer what the actual differences in configuration are.

A more sensitive display of differences in the configurations of two batches may be obtained by biplotting each batch separately and superimposing their h-configurations (the g scatters are of no interest in this context). These h-plots (Corsten and Gabriel, 1976) allow more detailed comparisons. Thus, for the two batches of 1973 storms, the two h-configurations are superimposed in Figure 15. Note that with a slight rotation

Figure 15: H-plots of Fast Storms (above) and Slow Storms (below)





four of the five pairs of $\underline{h}$-vectors can be made to overlap pretty well. The obvious exception is the $\underline{h}_R$ vector which is in almost opposite direction in the two configurations -- its correlations with the other variables must be of virtually opposite signs in the two batches. This agrees pretty well with Table 15.

We recapitulate the methods of comparing batches of multivariable data. To begin with, one should check the batch scatters to see whether they are reasonably elliptic in character. If there seem to be few outliers, long tails and/or strong concentration at one edge or corner (likely to be the zero point of the variables if the measurements are all non-negative) then the data should be readjusted in a manner similar to single batch data with such properties. If the variability seems to be systematically longer for the batch with larger means, transformations may be called for. A comparison of log (standard deviations) against log (mean) may show a fixed slope for some of the variables -- these variables' observations are likely to be more regularly scattered, i.e., have more equal variabilities, if they are re-expressed as

$$(\text{variable})^{1-\text{slope}} = (\text{re-expressed variable}) .$$

Such re-expression is a rough and ready method and the exponent should in general be rounded to the nearest 1/2. (Note that for a slope around 1, $(\cdot)^{1-\text{slope}}$ is to be read as $\log(\cdot)$). (Tukey, 1977, Chapters 3 and 4).

6.3. Comparison of three or more batches

When multivariate observations occur in several batches, or have been a priori classified into several categories, one may wish to compare these several sets of observations. Essentially, such comparisons of three or more batches are analogous to the comparison of two batches as described above. There are essentially two approaches to such comparisons: (1) comparing the several batches configurations of variables without reference to the location of the scatters, or, (2) comparing the several batches' unit-scatters against the background of one configuration of variables. These two approaches correspond to univariate comparisons of scale and location.

For comparisons (1) of configurations one would need separate variance-covariance configurations to be obtained and displayed for each batch. Thus, the $\underline{h}$-configuration of each batch would be obtained from its GH'-biplot. These configurations might then be displayed alongside one another and compared visually. If the number of variables is not very small, such visual comparisons may be quite difficult. Section 6.2 illustrated a comparison of two batches' 4-variate configurations. Consider how much more difficult the comparison of six batches would be if 10-variate configurations were displayed for each. Unfortunately we cannot suggest a simpler way of making such comparisons. They are complex and perhaps cannot be further simplified.

For location comparisons (2) one might begin by pooling all batches to obtain an overall estimate of the variables' configuration, i.e., the h-configuration in the GH'-biplot of all the data. One would then compare the batches by classifying the g-points according to the batches whose units they represent. And again, as in Section 6.1, one might summarize the scatter of each batch by a concentration ellipse. The comparison of batch scatters is then conveniently done by inspecting the locations and shapes of the different ellipses as in Section 6.1.

An alternative approach to location comparisons is to compare only the multivariate means of the several batches. A suitable metric for such comparisons is that of the "within batches" sums of squares of products (its use assumes that the variance-covariance configuration of the different batches are much the same). Thus, a biplot of the means of the several batches would show which batches differ from what other batches and on what variables these differences are evident.

Such an approach is analogous to MANOVA (multivariate analysis of variance). An application to meteorology has been studied in the context of the Israeli rainfall stimulation experiment (Gabriel, 1972).

6.4. Classification of new data into given categories -- Discriminant analysis

A common situation requires the classification of a new unit into one of several populations from which it might have originated. Thus, storms may be of a number of synoptic types and radar observations may be available for batches of earlier storms of each type. A new storm now occurs and one is asked to use its radar observations in order to allocate it to one synoptic type. Statistically, one would want to classify the new storm into the type whose batch's radar observations match the new storm's observations most closely. That essentially is the problem statisticians refer to as "*discrimination*" and the techniques they use go under the name of discriminant analysis. The subject is too large to explore here: Instead we refer the reader to Miller's (1964) monograph, written for meteorologists, to Lachenbruch's (1975) volume on discriminant analysis and to Gabriel and Pun's (1978) description of and program for two category discrimination by logistic techniques.

6.5 An Example - Different Techniques for Comparing Batches

To evaluate the three different ways of comparing several batches of multivariate data, we will study an example in some detail. We use historical data of annual precipitation in Illinois to simulate a weather modification situation. Suppose a "cloud seeding" operation had taken place in the years 1955 - 1960 in the Southern Illinois area, and that another such operation had been carried out during 1970-78 in the Northeastern part of Illinois. Also, suppose that no cloud seeding was carried out in Illinois at any other time or place. Central Illinois precipitation could, therefore, serve as concomitant observations to indicate "natural" precipitation; it would not have been "seeded" in either period. (The quotes are used since the data relate to simulated "operations", not to real ones).

To evaluate the effect of both "operations" it may be proposed to use 50 years' data, 1929-78, for the following five stations: Dubuque and Moline to represent Northeastern Illinois, "seeded" in Period IV -- 1970-78; St. Louis to represent Southern Illinois, "seeded" in Period II -- 1955-60; Peoria and Springfield to represent Central Illinois -- never "seeded". These 50 years also provide two "unseeded" periods for comparison, i.e., I - 1929-1954 and III - 1961-1969, as set out in Table 16 The corresponding data for annual precipitation are shown in Table 17 Note that these are actual precipitation data except that in the "operational" years each "target" station's precipitation was augmented to simulate effects of "seeding".

We are using simulated data for illustration because that allows us to anticipate the findings and then see how, and to what

Table 16 Areas and Periods of "Operations" and Comparisons

Period	No. of Years	Southern Illinois "Target" (St. Louis)	Northeastern Illinois "Target" (Dubuque, Moline)	Central Illinois Control (Peoria, Springfield)
I. 1929-54	26	Unseeded	Unseeded	Unseeded
II. 1955-60	6	"Seeded"	Unseeded	Unseeded
III. 1961-69	9	Unseeded	Unseeded	Unseeded
IV. 1970-78	9	Unseeded	"Seeded"	Unseeded
Total	50			

Table 17 Annual Precipitation at 5 Illinois Stations 1929-78

	DUB	MOL	PEO	SPR	STL
29	24.160	24.710	29.660	27.680	40.200
30	28.350	30.010	24.030	24.320	27.230
31	29.540	31.390	37.750	36.210	37.300
32	25.970	34.490	33.680	32.050	38.010
33	28.700	28.310	34.070	36.470	34.770
34	34.500	36.850	30.430	35.680	29.190
35	32.550	35.580	40.150	41.220	39.360
36	26.770	30.080	30.910	28.020	26.140
37	31.770	30.960	29.890	34.630	35.870
38	47.630	43.750	42.620	36.980	41.220
39	29.890	28.500	38.270	33.050	40.150
40	23.900	25.200	24.160	22.880	25.000
41	32.500	36.940	42.090	44.720	32.120
42	35.570	32.880	37.860	43.360	45.140
43	31.920	32.160	32.810	32.360	33.600
44	42.500	38.930	35.930	33.280	32.510
45	38.780	33.840	36.130	43.400	49.820
46	32.510	38.320	38.890	39.910	57.120
47	42.280	35.620	39.170	36.480	35.780
48	23.350	34.350	30.130	30.860	42.260
49	31.510	34.560	33.330	37.520	45.760
50	32.330	32.880	37.300	32.050	37.330
51	45.010	48.600	37.230	39.510	36.370
52	27.260	28.640	35.430	20.390	25.670
53	34.950	26.470	28.850	23.980	20.590
54	38.210	38.860	41.960	26.670	27.610
55	26.070	26.090	29.990	34.150	40.729
56	24.080	20.200	25.620	31.210	44.759
57	38.920	32.920	36.990	41.970	61.308
58	26.070	24.450	31.450	30.560	48.594
59	54.360	42.100	30.630	35.980	36.805
60	43.360	39.450	37.630	38.910	41.314
61	63.090	45.900	39.450	37.910	41.200
62	42.770	33.850	24.820	30.620	34.610
63	35.440	30.780	25.660	28.890	28.620
64	36.140	35.670	28.950	31.020	32.160
65	61.420	49.590	48.260	39.080	28.260
66	39.230	37.680	33.140	30.700	32.340
67	52.970	42.360	35.950	36.310	41.300
68	39.960	31.850	33.890	31.670	32.490
69	33.700	41.790	33.700	34.780	43.720
70	47.801	67.236	44.720	38.250	36.200
71	48.217	49.972	26.390	27.620	30.730
72	51.714	60.645	36.230	32.030	33.740
73	51.426	73.268	50.225	44.290	39.820
74	50.154	60.879	42.510	40.820	36.630
75	42.263	37.635	41.220	37.660	40.210
76	30.654	32.461	31.230	25.700	23.460
77	50.739	54.548	38.410	42.710	43.410
78	40.300	40.651	32.090	31.830	37.710

NOTE: These are actual precipitation data as obtained from the Illinois State Water Survey, except for the 1955-60 figures for St. Louis and the 1970-78 figures for Dubuque and Moline which are equal to 130% of the recorded natural precipitation.

extent, the analyses recover the simulated "effects". Thus, we should expect that during "operations" the precipitation at target stations would be higher and more variable.

We begin by examining the entire data set, irrespective of batches, i.e., operational or other years. Means, standard deviations, covariances and correlations are shown in Table 18, and the co-ordinates for the GH'-biplot are given in Table 19 -- the biplot having been fitted to residuals from the 5-variate centroid. This biplot is displayed in Figure 16.

Mean precipitation -- Table 18 -- is pretty uniform over the five Illinois stations -- perhaps a little lower in the center. Variability changes more strikingly, the standard deviations being appreciably lesser in Central Illinois. Correlations reflect the geographical location, the highest correlations being found for adjacent stations, i.e., Dubuque with Moline, Moline with Dubuque and to a lesser extent with Peoria, Peoria with Springfield and St. Louis with Springfield. Generally, correlation tapers off with distance between stations -- thus the St. Louis correlations with Dubuque and Moline are very low.

The biplot -- Figure 16 -- reflects this configuration of variation and covariation (since this GH'-biplot is mean-centered it conveys no information on means). The $\underline{h}$-arrow for the Central Illinois stations are shorter (less variability) than those for the stations in North and South Illinois. The order of the arrowheads reflects the geographical location of the five stations and so the angles subtended at the centroid are smaller for nearby stations and larger for far-away stations: Thus, the cosines decrease with distance, reflecting the decrease of correlations

Table 18 Measure of Location and Dispersion of the Entire Data Set

	Station				
	Dubuque	Moline	Peoria	Springfield	St. Louis
Means	38.111	37.897	35.043	34.503	37.507
Standard Deviation	9.626	10.805	6.013	5.474	8.203
Covariances / Correlations					
		77.643	24.937	18.916	1.022
	.7465		39.215	26.563	5.620
	.4308	.6036		22.878	13.405
	.3590	.4491	.6951		27.287
	.0129	.0634	.2717	.6077	

Table 19 GH'-biplot Co-ordinates for Entire Set

i	g_{i1}	g_{i2}	i	g_{i1}	g_{i2}	i	g_{i1}	g_{i2}
1929	-.083	-.193	1946	.014	-.338	1963	-.114	.154
30	-.182	.232	7	.023	.003	4	-.058	.095
1	-.085	-.068	8	-.070	-.058	5	.259	.161
2	-.104	-.046	9	-.051	-.163	6	-.014	.097
3	-.124	-.034	50	-.068	-.032	7	.130	-.007
4	-.048	.099	1	.133	.020	8	-.046	.077
5	-.017	-.114	2	-.149	.118	9	.005	-.095
6	-.158	.137	3	-.158	.271	70	.300	.057
7	-.102	-.007	4	.000	.156	1	.101	.178
8	+.129	-.048	5	-.154	-.094	2	.235	.156
9	-.106	-.094	6	-.234	-.145	3	.397	-.020
40	-.185	.226	7	.017	-.389	4	.269	.035
1	-.001	-.042	8	-.173	-.195	5	.051	-.067
2	-.014	-.193	9	.115	.078	6	-.127	.211
3	-.092	.032	60	.064	-.070	7	.227	-.064
4	.029	.071	1	.230	.020	8	.019	.031
5	.012	-.237	2	-.042	.098			

j	h_{j1}	h_{j2}
DUB	59.235	+14.807
MOL	71.587	+ 8.645
PEO	29.079	-12.736
SPR	22.569	-24.452
ST.L.	10.577	-54.246

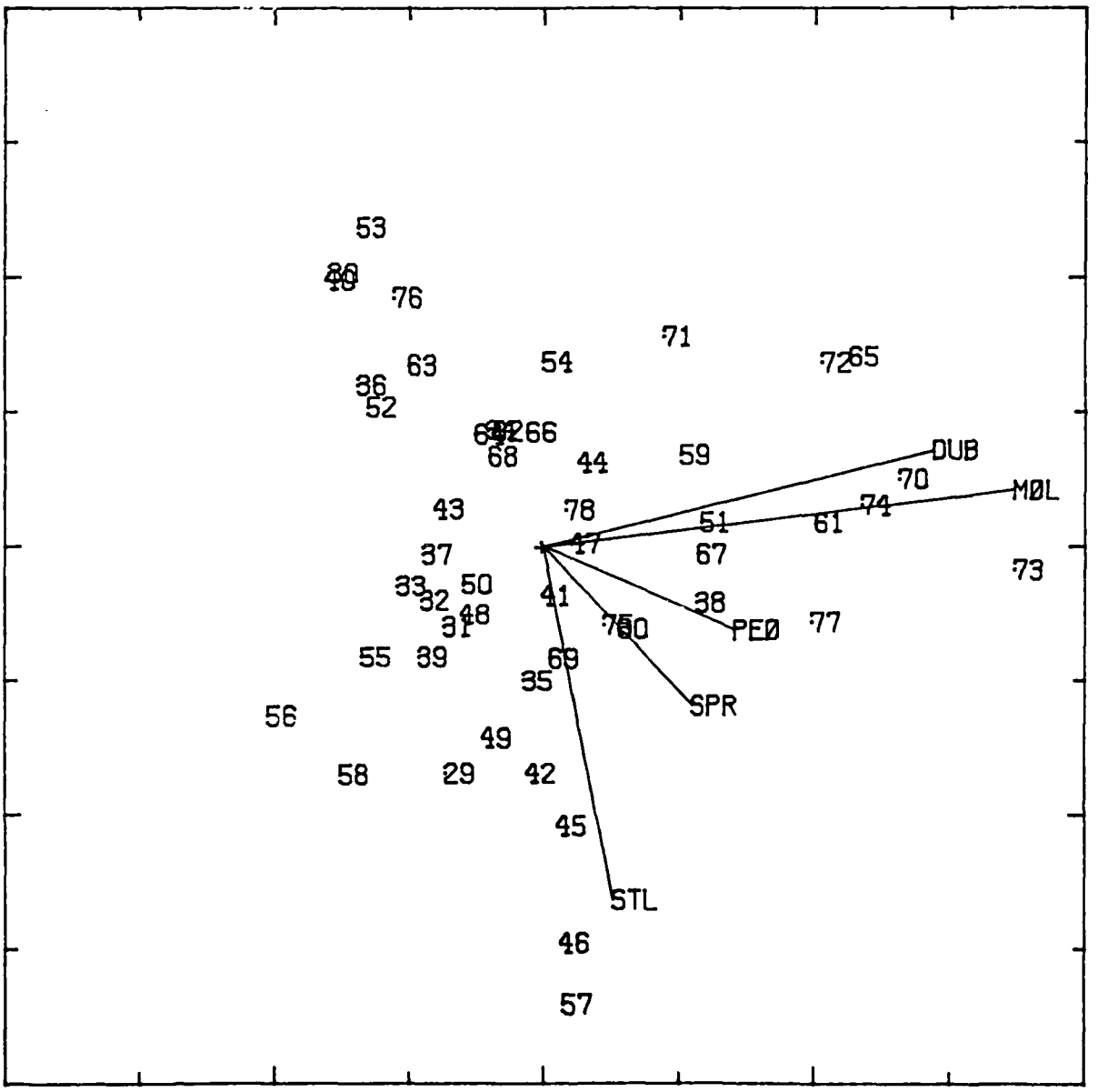
$\lambda_1 =$ 100.50

$\lambda_2 =$ 63.22

$\Sigma\lambda^2 =$ 16798.25

Goodness of Fit 0.8392

Figure 16: GH'-biplot of 50 Years' Illinois Rainfall (g-points Identified by Years; h-arrows by Station)



with distance. This biplot is therefore seen to provide a simple display of both the pattern of variation and the configuration of the correlations.

We next turn to the scatter of g-points in Figure 16, which displays the distribution of the 50 years about their 5-variate mean. This is a pretty evenly spread scatter -- no obvious outliers are evident, except perhaps 1957 at the bottom of the biplot. This shows unusually high precipitation in 1957 at St. Louis. (Note that this was a year in which St. Louis was a "seeding target"!). Indeed, on closer examination, we note 4 out of the 6 years "seeded" at St. Louis to have g-points pretty far out in the direction of the h-arrows for that station. Also, we see that 6 out of the 9 years of Northeast Illinois "seeding" have g-points far out in the biplot direction of the Dubuque and Moline h-arrows. This is suggestive of "seeding effects".

The distinction between the four periods may be accentuated by suppressing the dates on the biplot and substituting the number of the period, i.e., 1, 2, 3 or 4, at each g-point. This is done in Figure 17. This display emphasizes the predominance of g-points of periods II and IV in the directions of the h-arrows for, respectively, St. Louis and Dubuque/Moline.

Figure 18 is another version of this same GH'-biplot in which the individual years' g-points have been replaced by concentration ellipses for each of the periods. Now the comparisons are much easier to grasp. The average level of Period II precipitation is seen to be highest on the St. Louis target and that of Period IV on the Dubuque and Moline targets. The two unseeded periods, I and III, have fairly similar ellipses which

Figure 17: GH'-biplot of Illinois Rainfall (g-points Identified by Period)

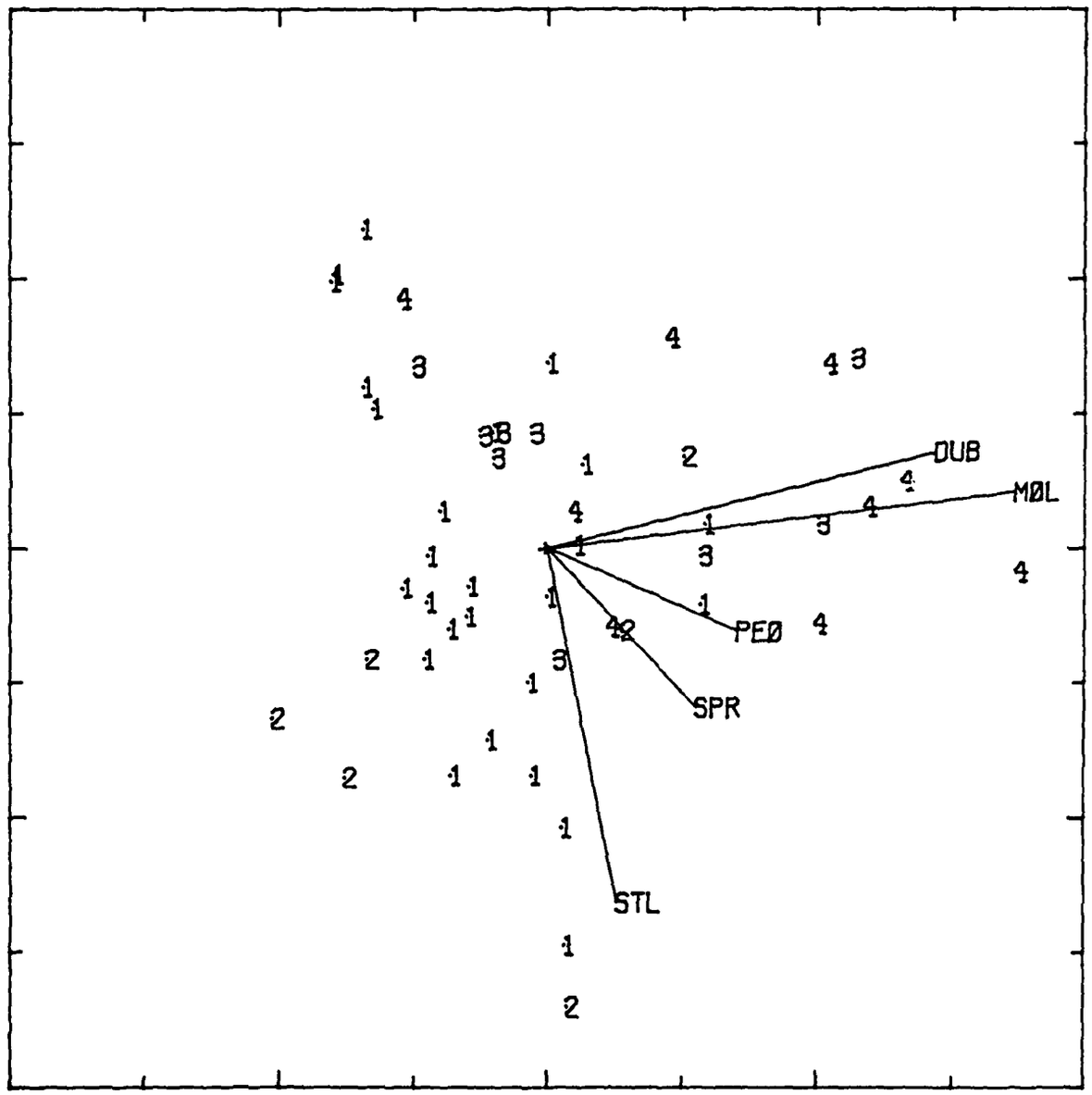
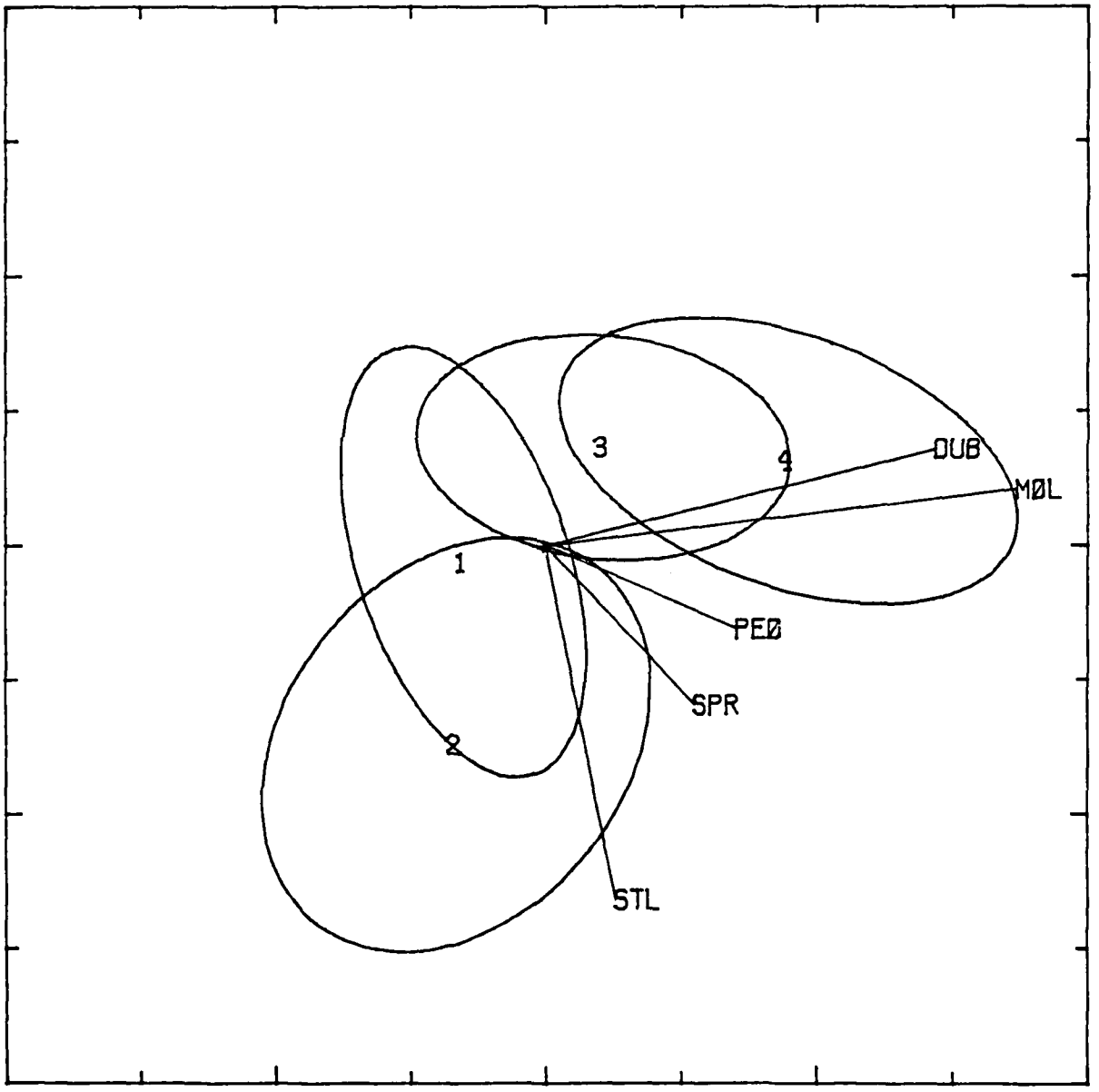


Figure 18: GH'-biplot of Illinois Rainfall With Concentration Ellipses for Periods



are not particularly high at any one of the stations.

Also note the different shapes of the ellipses, indicating differences in variability. The elongation of the Period IV ellipse along $\underline{h}_{DUB}$ and $\underline{h}_{MOL}$ suggests that the variance in Northeastern Illinois and the correlation between the stations must have been higher in the period when it was the "seeding target". Similarly, the ellipse for Period II is somewhat elongated along the direction of $\underline{h}_{STL}$. That indicates that when St. Louis was being "seeded" its variability was rather high.

Inspection of the GH' biplot of the entire data set has revealed differences in location as well as in variability and correlations. In most analyses this is likely to be the single most useful display. However, we will also illustrate the other two displays: The set of batch $\underline{h}$-plots which is designed specifically for comparisons of variability and correlation; and the MANOVA biplot which displays comparisons of means standardized for within batch variability.

For the comparison of periods, Table 20 gives the means, standard deviations, covariances and correlations and Table 21 the co-ordinates for the $\underline{h}$-plots of all periods. The four periods' $\underline{h}$-plots are shown together in Figure 19.

The $\underline{h}$-plots for the four periods -- Figure 19 -- look rather different at first glance, as do the standard deviations and correlations in Table 21. This is mostly due to the random variability between such small batches of data: It is well known that correlations based on samples of as few as 6 and 9 observations fluctuate wildly. Indeed, the comparison of Periods I and III which were both "unseeded" shows how large random variability

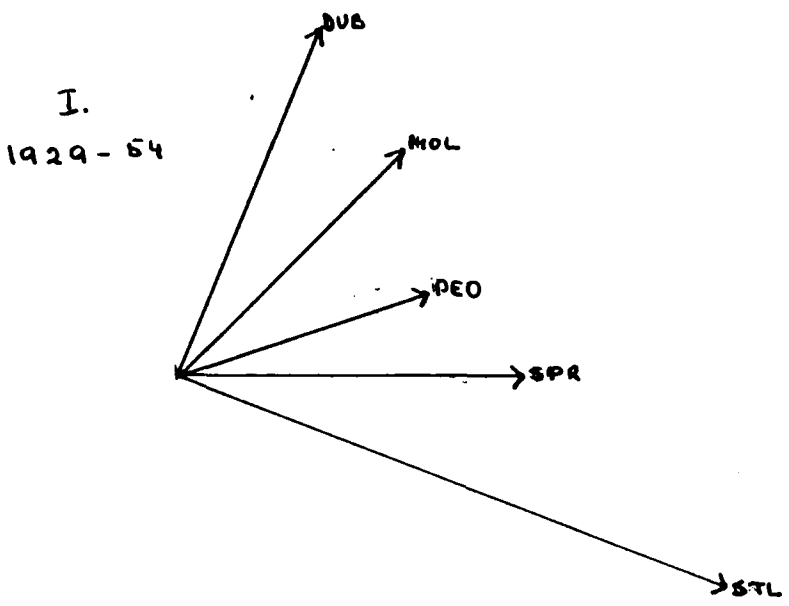
Table 20 Measures of Location and Dispersion of Four Periods of Year

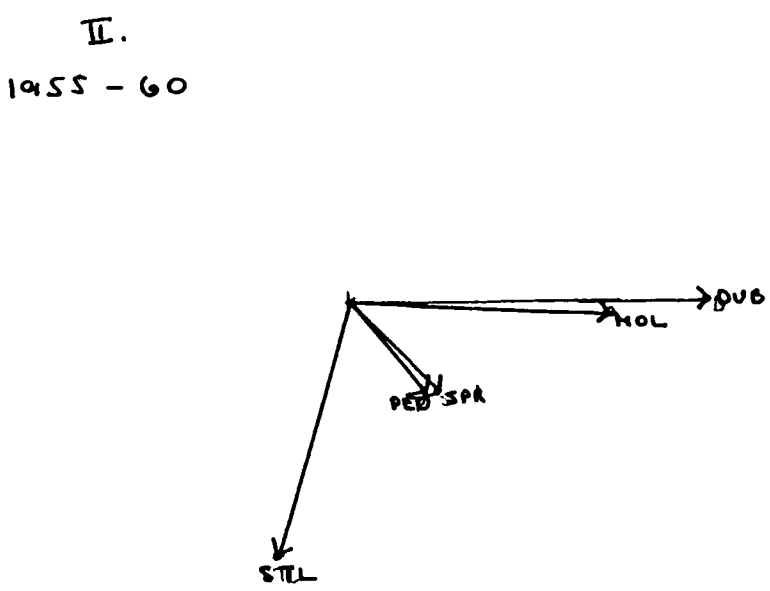
	Stations				
Means	DUB	MOL	PEO	SPR	ST.L
I. 1929-54	33.558	33.957	35.116	34.253	36.135
II. 1955-60	35.793	30.868	32.052	35.463	45.585
III. 1961-69	45.002	38.830	33.758	33.431	34.967
IV. 1970-78	45.919	53.033	38.112	35.657	36.123
Standard Deviations (diagonal), Covariances (above diagonal), Correlations (below diagonal)					
Period I	5.886	19.998	9.298	8.738	4.234
	.679	5.000	15.158	15.019	16.554
	.309	.593	5.115	19.073	21.562
	.250	.506	.628	5.934	35.309
	.083	.380	.484	.683	8.708
Period II	11.854	100.791	25.574	32.601	-25.972
	.973	8.739	24.873	26.371	-18.066
	.474	.626	4.550	16.728	15.964
	.620	.680	.829	4.435	15.637
	-.253	-.238	.405	.407	8.671
Period III	11.361	57.407	62.972	34.954	8.756
	.779	6.487	41.067	22.489	12.777
	.768	.877	7.215	23.575	2.621
	.842	.949	.895	3.652	8.416
	.135	.344	.063	.403	5.721
Period IV	6.999	81.964	23.608	29.373	26.095
	.840	13.946	70.461	60.135	32.450
	.453	.678	7.451	43.173	21.676
	.632	.649	.872	6.645	30.404
	.656	.410	.512	.805	5.682

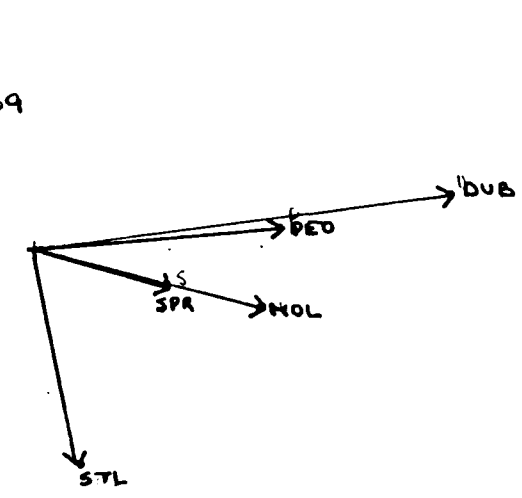
Table 21 h-plot Co-ordinates for the Four Periods

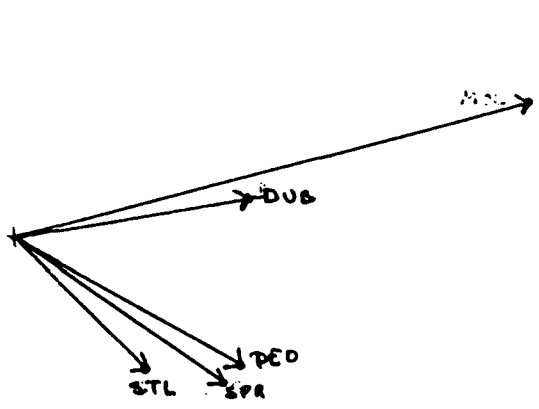
	Station					
Period I	DUB	MOL	PEO	SPR	STL	
h_{j1}	10.9	17.0	18.5	25.4	39.2	$\lambda_1 = 54.1$
h_{j2}	24.9	15.9	5.6	- 0.5	-16.1	$\lambda_2 = 34.1$
			Goodness of Fit = .8236			
Period II						
h_{j1}	26.3	19.4	5.5	6.4	- 5.3	$\lambda_1 = 34.2$
h_{j2}	- 0.1	- 0.9	- 6.6	- 6.5	-18.4	$\lambda_2 = 20.7$
			Goodness of Fit = .9596			
Period III						
h_{j1}	30.9	16.7	18.3	9.8	3.5	$\lambda_1 = 40.9$
h_{j2}	3.8	- 4.1	1.6	- 2.4	-15.5	$\lambda_2 = 16.7$
			Goodness of Fit = .9072			
Period IV						
h_{j1}	17.2	38.2	16.6	15.2	9.6	$\lambda_1 = 48.5$
h_{j2}	2.7	9.5	- 9.4	-10.6	- 9.4	$\lambda_2 = 19.6$
			Goodness of Fit = .9120			

Figure 19: Illinois Rainfall; h-plots for Each of the Four Periods









6/29

really is: Note the Dubuque-Springfield correlation being 0.250 in Period I versus 0.842 in Period III! This illustration should serve as a warning against drawing far reaching conclusions about variability and correlation on the basis of small data sets.

Despite the smallness of the samples, there is some consistency in the four $\underline{h}$-plots of Figure 19. The geographical gradient from Northeastern through Central to Southern Illinois is shown consistently in all periods except III in which there is one inversion in the geographical order -- between Moline and Peoria. The orientation of the geographical gradient changes from period to period, but the gradient persists, illustrating that some general patterns may be revealed even from small samples of data.

It is difficult to find the expected "effects of seeding" in these displays. "Target" variability was expected to increase during "seeding" -- the Moline $\underline{h}$-arrow is unusually long in Period IV. But the Dubuque $\underline{h}$-arrow is rather short in Period IV and the St. Louis $\underline{h}$-arrow is not particularly long in Period II. Nor does the angle between $\underline{h}_{DUB}$ and $\underline{h}_{MOL}$ seem unusually low in Period IV -- as it should have been if "seeding" had increased correlation between "target" stations. Indeed, if we check back to Table 20 we note that these "expected effects" did not occur. It is not $\underline{h}$-plot display that obscured them, but the magnitude of random fluctuations in small samples.

Table 22 Estimate of "Error" Variance Based on Two Periods Without Operations

Stations	STATIONS				
	DUB	MOL	PEO	SPR	STL
DUB	7.585	29.067	22.310	15.094	5.330
MOL	.692	5.541	21.439	16.830	15.638
PEO	.516	.679	5.696	20.165	16.970
SPR	.364	.555	.647	5.469	28.789
STL	.087	.349	.368	.651	8.086

NOTE: Covariances above diagonal, standard deviations in diagonal, correlations below diagonal

Finally, we turn to the comparison of means -- shown above in Table 20 -- as standardized by random variation and covariance That is the multivariate analysis of variance (MANOVA) approach. Standardization is effected by weighting with the inverse of an estimated variance-covariance. The usual estimate is the "within" matrix of variances and covariances. (In this example it would be estimated by pooling the bottom four panels of Table 20 with weights 25, 5 and 8, respectively to a total of 46 degrees of freedom for error), but in the present instance we prefer to pool only the two "unseeded" periods so as to avoid possible contamination of the estimate by "seeding" effect. Thus, we pooled panels I and III of Table 20 with weights 25 and 8, respectively, yielding 33 error degrees of freedom -- Table 22.

The MANOVA calculations are shown in Table 23 and the corresponding JK'-biplot of the four period means at the five stations is displayed in Figure 20. Each period mean is surrounded by a "comparison circle" which gives an idea of the random variability of each of those period means. (The method of calculation of the radiuses of these circles is shown in Table 23; for a discussion of the rationale of these methods see Gabriel, 1972). The interpretation of these circles is simple. Any two periods whose circles intersect do not differ more than expected by chance; any two periods whose circles are disjoint differ significantly, i.e., more than expected by chance. In this application chance variability is read as 95% of random variability overall; thus a 5% chance -- level of significance -- is allowed of finding significance on some pair of periods that does not really differ. (Other levels could be chosen, e.g., for a 1%

Table 23 Calculations for MANOVA and JK'-biplot of Means

X: Batch Means

	DUB	MOL	PEO	SPR	ST.L
I	33.558	33.957	35.116	34.253	36.135
II	35.793	30.868	32.052	35.463	45.585
III	45.002	38.830	33.758	33.431	34.967
IV	45.919	53.033	38.112	35.657	36.123

S^{-1} = Inverse of Error Variances

	DUB	MOL	PEO	SPR	ST.L
D	.036 037	-.032 422	-.004 803	-.004 409	.008 005
M	-.032 422	.092 402	-.029 702	-.007 462	-.008 463
P	-.004 803	-.029 702	.072 456	-.032 675	.003 077
S	-.004 409	-.007 462	-.032 675	.089 663	-.028 855
L	.008 005	-.008 463	.003 077	-.028 855	.028 573

N = Diagonal Matrix of Sample Sizes

i	I	II	III	IV
n_i	26	6	9	9

$X'NXS^{-1}$

	DUB	MOL	PEO	SPR	ST.L
DUB	-1.475 78	102.070 99	-46.893 78	-16.250 12	-8.097 50
MOL	-35.447 94	188.119 76	-57.863 27	-21.508 86	-23.885 08
PEO	-13.071 80	39.692 80	-6.663 76	-2.790 28	-7.861 94
SPR	-2.942 10	9.749 88	-2.943 06	-1.247 80	0.224 44
STL	10.422 75	-31.365 39	1.239 63	.365 76	13.901 73

First Two Eigenvectors

	DUB	MOL	PEO	SPR	ST.L
$\underline{w}_1'$	-2.42	-4.16	-0.84	-0.19	0.72
$\underline{w}_2'$	-4.62	-0.81	1.26	0.06	-1.71
Column Markers:	$\underline{k}_D$	$\underline{k}_M$	$\underline{k}_P$	$\underline{k}_S$	$\underline{k}_L$

(These eigenvectors are standardized so that $\underline{w}_1 S^{-1} \underline{w}_1 = \underline{w}_2' S^{-1} \underline{w}_2 = 1$, $\underline{w}' S^{-1} \underline{w}_2 = 0$).

$J = XS^{-1}(\underline{w}_1, \underline{w}_2)$ Row Markers

I: $j_1' = (.822, .597)$

II: $j_2' = (1.990, -1.002)$

III: $j_3' = (-.087, -1.108)$

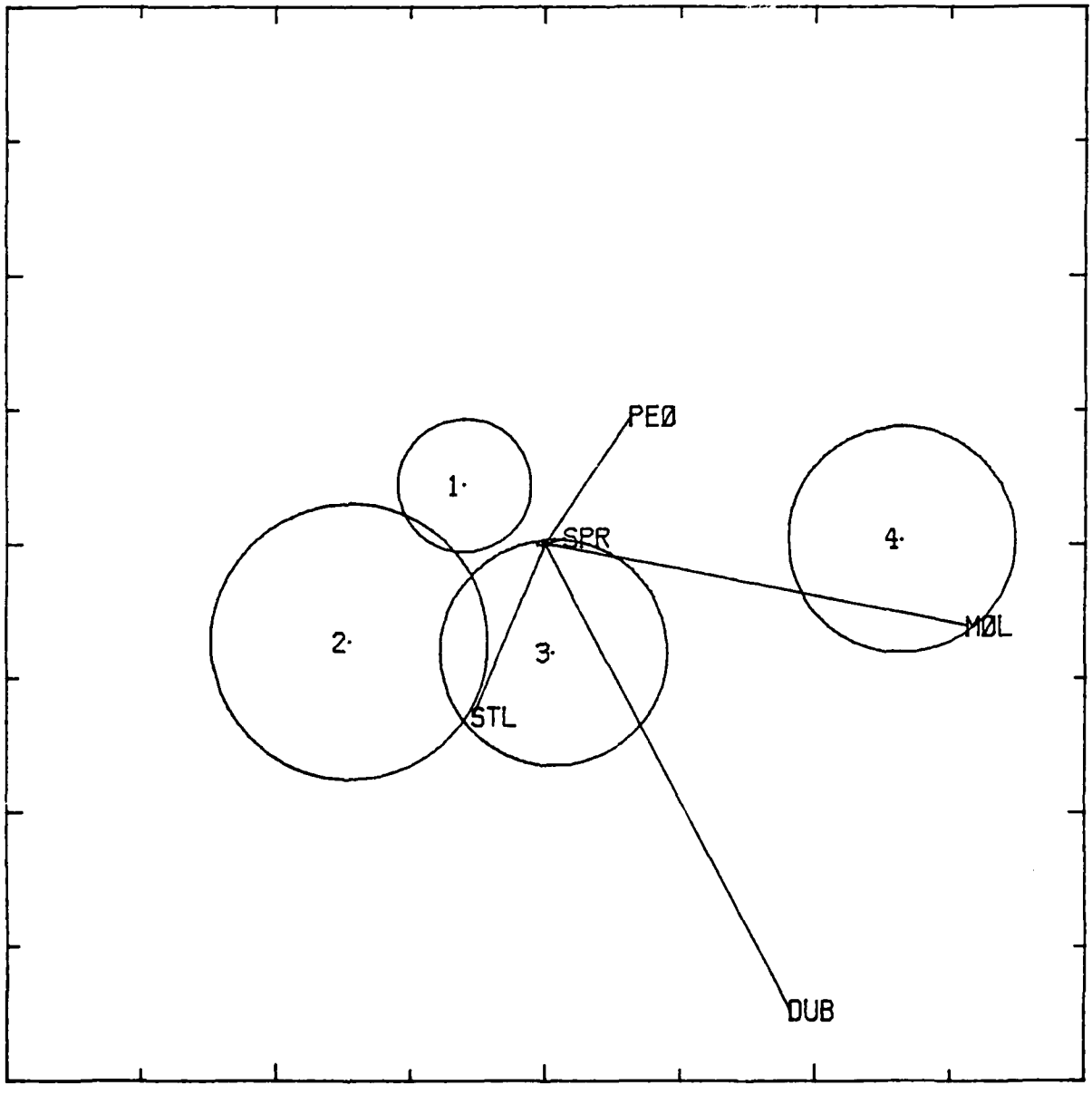
IV: $j_4' = (-3.615, 0.051)$

θ: Critical Value $\theta = 0.420$ $330/(1-\theta) = 23.892$

This is the upper 5% point of the maximum characteristic root distribution for 5 variables, 4 samples and 33 d.f for error Heck, 1960).

i	I	II	III	IV
Radius of Comparison Circle $\sqrt{[330/(1-\theta)2n_i]}$	.679	1.411	1.152	1.152

Figure 20: JK'-biplot of Period Means of Illinois Rainfall (With 5% Comparison Circles)



level the circles would be larger -- because a larger θ would be read from Heck's charts -- and fewer differences found significant. That would be a safer, but less revealing, strategy).

The comparison circle significance tests on Figure 20 show Period IV's means to differ significantly from the other three periods' means. Periods I and III barely differ, and Period II does not quite differ significantly from either of these.

The scatter of the four period means -- j-points -- can be related to the configuration of the five station measurements -- h-arrows. It is evident from Figure 20 that Period IV had large means in Northeastern Illinois, especially in Moline and less so at Dubuque. This confirms the "effect of seeding" in that area in Period IV, though the difference between Moline and Dubuque is unexpected. The small, and non-significant, difference between Period II and the "unseeded" Periods I and III is mostly in the direction of $\underline{k}_{STL}$, indicating higher precipitation at St. Louis in that period -- which is as it should be since that was where "seeding" was carried out. The other small, though significant, difference is between the two unseeded periods, I and III; it is not quite clear what this is due to and it may well be a "Type I error", i.e., a falsely significant finding when no true difference exists.

It is evident that much the same general picture was obtained from the comparison of means on the MANOVA JK'-biplot of Figure 20 and from the comparison of scatters (ellipses of concentration) in the GH'-biplot of Figure 18. Indeed, both these biplots are projections of the data matrix, with the four batches of points and five columns, onto different two-dimensional planes. The GH'-biplot describes the entire variability of the data, whereas the JK'-biplot shows only the scatter of means. The latter therefore emphasizes the <u>differences</u> between the periods rather than what they have in

common. To the extent that the latter configuration is different from the former, it is because of this different emphasis.

It must be remarked that the approximate significance tests illustrated in Figure 20 are valid only to the extent that the required assumptions are satisfied. These are (1) multi-normal distribution of precipitation in the five stations, (2) equal variances-covariances, (3) independence of observations. For annual precipitation data (1) may be a reasonable approximation and (2) would probably hold pretty well unless seeding effects were large. Whether successive annual precipitation amounts are independent is more doubtful -- though a recent study (Gabriel and Petrondas, in preparation) suggests that assumption (3) is not seriously wrong. If it were not tenable, then it would be wrong to regard the four periods as random samples and the significance tests would be quite invalid. That is a crucial point in many meteorological applications; it is often doubtful whether successive observations can be considered independent and thus the application of significance test is suspect. The emphasis in this chapter was therefore on exploratory data analysis rather than on significance testing of hypotheses -- that seems to be of more use in meteorological research.

7. ON TESTS OF SIGNIFICANCE

7.1. The logic of significance testing

A test of significance provides a decision on whether to regard an observed phenomenon as "random" or real. In other words, could the phenomenon have arisen in a manner analogous to the outcome of a game of chance, or does it reflect a real pattern? The issue of randomness versus real effects is often of great importance: Are there real periodicities in precipitation, trends in temperature, etc., and could the claimed effect of cloud seeding programs be merely due to chance? The use of significance tests to resolve these questions is, however, not as straightforward as might be thought. A few words on this topic are in place.

Significance tests are designed to disentangle real from random effects. They do so by checking whether the observations seem "non-random" in the direction in which real effects are a priori thought likely to occur. Thus, when a cloud seeding experiment is designed, the hypothesis of no effect is to be tested against that of augmented precipitation subsequent to seeding. When this expected effect is precisely defined in terms of location of precipitation, time, method of measurement, etc., a significance test can properly be applied.

Significance tests are of more doubtful validity when they are applied to "effects" which were first observed during the experiment itself. For example, the Swiss Grossversuch III was designed to reduce hail but observation of increased rainfall led to significance testing of augmented precipitation. It is a common occurrence that apparent effects are first observed in a particular area or at a particular time, e.g., after some change in seeding protocol, and then these particular "effects" are tested for significance. The validity of such testing is often in doubt because it does not take into account the fact that the one most striking phenomenon observed on the data was singled out for testing. A multiplicity of other phenomena were not tested because they did not happen to occur in such a striking form on those particular data. Significance tests are not usually designed to accomodate such selection of effects for testing. When it is done, the multiplicity of possible choices dilutes the significance and leads to spuriously "significant" results.

When non-experimental data are tested for significance, one should have even greater concern for the validity of inferences. Why was a particular phenomenon chosen for testing? Surely because it was observed to be remarkable. If so, the results of significance tests are strongly biased in favor of deciding on non-randomness. Tests would be valid only if carried out on new data sets, independent of those which suggested the phenomenon.

A convenient terminology is that distinguishing confirmatory from exploratory analyses (Tukey, 1977). The latter are essentially inductive, sifting through data for leads, patterns, suggestions and ideas. The former are of a more deductive and rigidly defined character -- they follow a protocol laid out in advance for the confirmation or refutation of a particular issue -- as in the prior hypothesis on precipitation to be confirmed or rejected by a cloud seeding experiment (Gabriel, 1980).

No doubt there is much more exploration than confirmation in scientific work, especially in non-laboratory situations. And these are common in meteorology. Application of significance tests in exploratory analyses cannot be regarded as a rigorous, well defined procedure: At best it serves to give vague indications of the relative roles of randomness and real effects.

7.2. The exploratory nature of multivariate analysis

Multivariate analysis, be definition, deals with a multiplicity of measures, none of which has been identified as the unique or principal bearer of the information sought. If a problem were closely defined and circumscribed, a single variable or function of variables would have been likely to emerge as the measure most relevant to the problem at hand. The analysis then would have lost its multivariate character. The simultaneous study of several variables thus implies that the subject is not narrowly focused and a definite hypothesis about the phenomena under study has not yet emerged. Hence multivariate analyses are unlikely to be confirmatory. Conversely, a confirmatory study is most likely to be univariate; the topic to be tested has been formulated precisely and allows confirmation. Exploratory studies are often multivariate, and allow the investigator to search for effects among a multiplicity of variables.

7.3. Significance tests in multivariate analysis

We have argued that multivariate analysis is mostly exploratory, and that exploratory studies do not in general depend much on significance tests. Hence the role of significance testing in multivariate analysis is likely to be minimal. This chapter has therefore not stressed topics of significance testing. Readers who still wish to apply tests of significance to multivariate data are referred to Morrison's (1976) excellent elementary text and to Essenwanger's (1976) more advanced volume. They will find tests for the types of comparisons discussed in Section 6 as well as for other types of multivariate analyses of data from Gaussian distributions. For a description of methods which are more robust against non-normality, readers are referred to Gnanadesikan (1977). The present author hopes that the convenience of a single summary or significance level will not deter his readers from exploring their data. He also hopes that the present chapter may help his readers to look at their data and discover what they have to tell.

ACKNOWLEDGMENTS

Computations were carried out by means of program BIPLOT (available from the author) and plotting package BGRAPH (designed by M. C. Tsianco) Mike Tsianco, David Gheva, and Sandra Plumb's help with the examples and computations is gratefully acknowledged, as is Marie Butler's typing.

REFERENCES

Bradu, D. and Gabriel, K. R. (1978). "The biplot as a diagnostic tool for models of two-way tables." Technometrics, 20, 47-68.

Corston, L. C. A. and Gabriel, K. R. (1976). "Graphical exploration in comparing variance matrices." Biometrics, 32, 851-863.

Essenwanger, O. (1976). Applied Statics in Atmospheric Science. Amsterdam: Elsevier.

Gabriel, K. R. (1971). "The biplot - graphic display of matrices with application to principal component analysis." Biometrika, 58, 453-467.

Gabriel, K. R. (1972). "Analysis of metrorological data by means of canonical decomposition and biplots." Journal of Applied Meteorology, 11, 1071-1077.

Gabriel, K. R. (1978). "The complex correlational biplot." In Theory Construction and Data Analysis in the Behavioral Sciences. Shye, S. (ed.). San Francisco: Jossey-Bass, 350-370.

Gabriel, K. R. (1981). "On the role of physicists and statisticians in weather modification experimentation." Bulletin of American Meteorological Society, 62.

Gabriel, K. R. and Pun, C. F. (1978). "Binary prediction of weather events with several predictors." Technical Report submitted to National Weather Service.

Gabriel, K. R. and Tsianco, M. (1980). "Diagnosis and fit of a harmonic model to meteorological data. (In preparation).

Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. New York: Wiley.

Gnanadesikan, R. and Kettenring, J. R. (1972). "Robust estimates, residuals and outlier detection with multiresponse data." Biometrics, 28, 81-124.

Guttman, L. (1954). "A new approach to factor analysis: The Radex." In Mathematical Thinking in the Social Sciences. Lazarsfeld, P. (ed.). New York: Free Press.

Lachenbruch, P.A. (1975). Discriminant Analysis. New York: Hafner.

Miller, R.G. (1964). "Regression estimation of event probabilities." Technical Report No. 1. Contract CWB-10704, The Travelers Research Center, Inc., Hartford, Conn., 153 pp.

Morrison, D.F. (1976). Multivariate Statistical Methods (2nd edition). New York: McGraw-Hill.

Pielke, R.A. and Biondini, R. (1977). "Rainfall in the EML target area as a function of synoptic parameter." Unpublished report.

Ramsey, A.K., Shepard, R.N. and Nerlove, S.B. (1972). Multidimensional Scaling, Vols. I and II. New York: Seminar Press.

Tukey, J.W. (1977). Exploratory Data Analysis. Reading Mass.: Addison-Wesley.

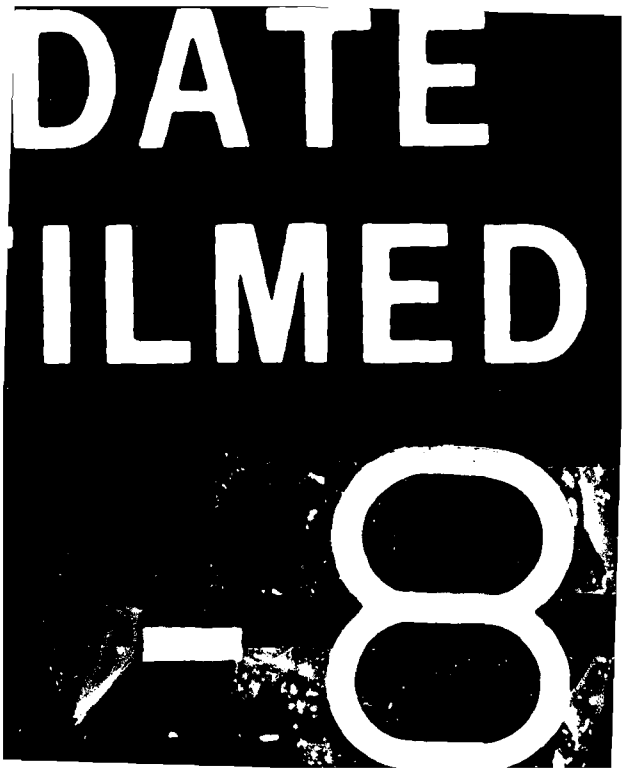
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A096 400 / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) Exploratory Multivariate Analysis: a graphical approach	5. TYPE OF REPORT & PERIOD COVERED (9) Technical Report / 6. PERFORMING ORG. REPORT NUMBER (14) TR-81/1
7. AUTHOR(s) (10) K. Ruben Gabriel	8. CONTRACT OR GRANT NUMBER(s) (15) N 00014-80-C-0387
9. PERFORMING ORGANIZATION NAME AND ADDRESS Division of Biostatistics, University of Rochester Medical School Rochester, NY 14642	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, Virginia 22217 (12) 137	12. REPORT DATE (11) January, 1981 / 13. NUMBER OF PAGES 134
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified / 15a. DECLASSIFICATION/ DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.	
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)	
18. SUPPLEMENTARY NOTES	
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Multivariate analysis; biplot; principal components; factor analysis; cluster analysis; Hotelling's T^2; MANOVA; discriminant analysis	
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) This is an introduction and commentary on multivariate statistical methods which leans heavily on the biplot. The latter is used as a didactical device to clarify the structure of multivariate methods of exploratory data analysis. Topics dealt with are explorations of single batches in terms of variances and correlations and distances between subjects. Principal components, clustering methods and factor analysis are	

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

307200

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


DATE
ILMED
-8